

# COMMUNITY DETECTION IN RANDOM NETWORKS

Ery Arias-Castro<sup>1</sup> and Nicolas Verzelen<sup>2</sup>

We formalize the problem of detecting a community in a network into testing whether in a given (random) graph there is a subgraph that is unusually dense. We observe an undirected and unweighted graph on  $N$  nodes. Under the null hypothesis, the graph is a realization of an Erdős-Rényi graph with probability  $p_0$ . Under the (composite) alternative, there is a subgraph of  $n$  nodes where the probability of connection is  $p_1 > p_0$ . We derive a detection lower bound for detecting such a subgraph in terms of  $N, n, p_0, p_1$  and exhibit a test that achieves that lower bound. We do this both when  $p_0$  is known and unknown. We also consider the problem of testing in polynomial-time. As an aside, we consider the problem of detecting a clique, which is intimately related to the planted clique problem. Our focus in this paper is in the quasi-normal regime where  $np_0$  is either bounded away from zero, or tends to zero slowly.

**Keywords:** community detection, detecting a dense subgraph, minimax hypothesis testing, Erdős-Rényi random graph, scan statistic, planted clique problem, sparse eigenvalue problem.

*Dedicated to the memory of Yuri I. Ingster*

## 1 Introduction

In recent years, the problem of detecting communities in networks has received a large amount of attention, with important applications in the social and biological sciences, among others (Fortunato, 2010). The vast majority of this expansive literature focuses on developing realistic models of (random) networks (Albert and Barabási, 2002; Barabási and Albert, 1999), on designing methods for extracting communities from such networks (Girvan and Newman, 2002; Newman, 2006; Reichardt and Bornholdt, 2006) and on fitting models to network data (Bickel et al., 2011).

The underlying model is that of graph  $\mathcal{G} = (\mathcal{E}, \mathcal{V})$ , where  $\mathcal{E}$  is the set of edges and  $\mathcal{V}$  is the set of nodes. For example, in a social network, a node would represent an individual and an edge between two nodes would symbolize a friendship or kinship of some sort shared by these two individuals. In the literature just mentioned, almost all the methodology has concentrated on devising graph partitioning methods, with the end goal of clustering the nodes in  $\mathcal{V}$  into groups with strong inner-connectivity and weak inter-connectivity (Bickel and Chen, 2009; Lancichinetti and Fortunato, 2009; Newman and Girvan, 2004).

In this euphoria, perhaps the most basic problem of actually detecting the *presence* of a community in an otherwise homogeneous network has been overlooked. From a practical standpoint, this sort of problem could arise in a dynamic setting where a network is growing over time and monitored for clustering. From a mathematical perspective, probing the limits of detection (i.e., hypothesis testing) often offers insight into what is possible in terms of extraction (i.e., estimation).

Many existing community extraction methods can be turned into community detection procedures. For example, one could decide that a community is present in the network if the modularity of Newman and Girvan (2004) exceeds a given threshold. To set this threshold, one needs to define a null model. Newman and Girvan (2004) implicitly assume a random graph conditional on

<sup>1</sup>Department of Mathematics, University of California, San Diego, USA

<sup>2</sup>INRA, UMR 729 MISTEA, F-34060 Montpellier, FRANCE

the node degrees. Here, we make the simplest assumption that the null model is an Erdős-Rényi random graph (Bollobás, 2001).

In this context, we also touch on another line of work, that of detecting a clique in a random graph — the so-called Planted (or Hidden) Clique Problem (Alon et al., 1998; Dekel et al., 2011; Feige and Ron, 2010). Although the emphasis there is to find the detection performance of computationally tractable algorithms, we mostly ignore computational consideration and simply establish the absolute detection limits of any algorithm whatsoever.

## 1.1 The framework

We address a stylized community detection problem, where the task is to detect the presence of clustering in the network and is formalized as a hypothesis testing problem. We observe an *undirected* graph  $\mathcal{G} = (\mathcal{E}, \mathcal{V})$  with  $N := |\mathcal{V}|$  nodes. Without loss of generality, we take  $\mathcal{V} = [N] := \{1, \dots, N\}$ . The corresponding adjacency matrix is denoted  $\mathbf{W} \in \{0, 1\}^{N \times N}$ , where  $W_{i,j} = 1$  if, and only if,  $(i, j) \in \mathcal{E}$ , meaning there is an edge between nodes  $i, j \in \mathcal{V}$ . Note that  $\mathbf{W}$  is symmetric, and we assume that  $W_{ii} = 0$  for all  $i$ . Under the null hypothesis, the graph  $\mathcal{G}$  is a realization of  $\mathbb{G}(N, p_0)$ , the Erdős-Rényi random graph on  $N$  nodes with probability of connection  $p_0 \in (0, 1)$ ; equivalently, the upper diagonal entries of  $\mathbf{W}$  are independent and identically distributed with  $\mathbb{P}(W_{i,j} = 1) = p_0$  for any  $i \neq j$ . Under the alternative, there is a subset of nodes indexed by  $S \subset \mathcal{V}$  such that  $\mathbb{P}(W_{i,j} = 1) = p_1$  for any  $i, j \in S$  with  $i \neq j$ , with everything else the same. We assume that  $p_1 > p_0$ , implying that the connectivity is stronger between nodes in  $S$ . When  $p_1 = 1$ , the subgraph with node set  $S$  is a clique. The subset  $S$  is not known, although in most of the paper we assume that its size  $n := |S|$  is known.

We study detectability in this framework in asymptotic regimes where  $n, N \rightarrow \infty$ , and  $p_0, p_1$  may also change; all these parameters are assumed to be functions of  $N$ . A test  $T$  is a function that takes  $\mathbf{W}$  as input and returns  $T = 1$  to claim there is a community in the network, and  $T = 0$  otherwise. The (worst-case) risk of a test  $T$  is defined as

$$\gamma_N(T) = \mathbb{P}_0(T = 1) + \max_{|S|=n} \mathbb{P}_S(T = 0),$$

where  $\mathbb{P}_0$  is the distribution under the null and  $\mathbb{P}_S$  is the distribution under the alternative where  $S$  indexes the community. We say that a sequence of tests  $(T_N)$  for a sequence of problems  $(\mathbf{W}_N)$  is asymptotically powerful (resp. powerless) if  $\gamma_N(T_N) \rightarrow 0$  (resp.  $\rightarrow 1$ ). Practically speaking, a sequence of tests is asymptotically powerless if it does not perform substantially better than any guessing that ignores the adjacency matrix  $\mathbf{W}$ . We will often speak of a test being powerful or powerless when in fact referring to a sequence of tests and its asymptotic power properties.

## 1.2 Closely related work

We take the beaten path, following the standard approach in statistics for analyzing such composite hypothesis testing problems, in particular, the work of Ingster (1997) and others (Donoho and Jin, 2004; Hall and Jin, 2010; Ingster and Suslina, 2002) on the detection of a sparse (normal) mean vector. Most closely related to our work is that of Butucea and Ingster (2011). Specializing their results to our setting, they derive lower bounds and upper bounds for the same detection problem when the graph is directed and the probability of connection under the null (denoted  $p_0$ ) is fixed, which is a situation where the graph is extremely dense. Their work leaves out the interesting regime where  $p_0 \rightarrow 0$ , which leads to a null model that is much more sparse.

### 1.3 Main Contribution

Our main contribution in this paper is to derive a sharp detection boundary for the problem of detecting a community in a network as described above. We focus here on the quasi-normal regime<sup>3</sup> where  $np_0$  is either bounded away from zero, or tends to zero slowly, specifically,

$$\log \left( 1 \vee \frac{1}{np_0} \right) = o \left[ \log \left( \frac{N}{n} \right) \right]. \quad (1)$$

On the one hand, we derive an information theoretic bound that applies to all tests, meaning conditions under which all tests are powerless. On the other hand, we display a test that basically achieves the best performance possible. The test is the combination of the two natural tests that arise in [Butucea and Ingster \(2011\)](#) and much of the work in that field ([Arias-Castro et al., 2011](#); [Ingster et al., 2010](#)):

- *Total degree test.* This test rejects when the total number of edges is unusually large. This is global in nature in that it cannot be directly turned into a method for extraction.
- *Scan (or maximum modularity) test.* This test amounts to turning modularity into a test statistic by rejecting when its maximum value is unusually large. It is strictly speaking the generalized likelihood ratio test under our framework.

We also consider the situation, common in practice, where  $p_0$  is unknown. Interestingly, the detection boundary becomes larger than in the former setting when  $n$  is moderately sparse. We derive the corresponding lower bound in this situation and design a test that achieves this bound. The test is again the combination of the two tests:

- *Degree variance test.* This test is based on the differences between two estimates for the degree variance, an analysis of variance of sorts. (Note that the total degree test cannot be calibrated without knowledge of  $p_0$ .)
- *Scan test.* This test can be calibrated in various ways when  $p_0$  is unknown, for example by estimation of  $p_0$  based on the whole graph, or by permutation. We study the former.

Finally, we consider various polynomial-time algorithms, the main one being a convex relaxation of the scan test based on a sparse eigenvalue problem formulation. Our inspiration there comes from the recent work of [Berthet and Rigollet \(2012\)](#). We discuss the discrepancy between the performances of the scan test and the relaxed scan test and compare it with other polynomial-time tests.

We summarize our findings in [Tables 1 and 2](#), where

$$R = \frac{\sqrt{n}(p_1 - p_0)}{\sqrt{p_0(1 - p_0)}}$$

is (up to  $\sqrt{n/2}$  factor) the SNR for detecting the dense subgraph when it is known.

---

<sup>3</sup>The quasi-Poisson regime where  $np_0 \rightarrow 0$  polynomially fast is qualitatively different and necessitates different proof arguments. This is beyond the scope of this paper and will appear somewhere else.

Table 1: Detection boundary and near-optimal algorithms. For any sequence  $a$  and  $b$  going to infinity,  $a \lll b$  (resp.  $a \ggg b$ ) means that there exists  $\epsilon > 0$  arbitrarily small such that  $a \leq b^{1-\epsilon}$  (resp.  $a \geq b^{1+\epsilon}$ )

	$p_0$ known		$p_0$ unknown	
	$n \lll N^{2/3}$	$n \ggg N^{2/3}$	$n \lll N^{3/4}$	$n \ggg N^{3/4}$
$p_0 \gg \frac{\log(N/n)}{n}$	$R > 2\sqrt{\log(N/n)}$	$R > N/n^{3/2}$	$R > 2\sqrt{\log(N/n)}$	$R > N^{3/4}/n$
$p_0 \ll \frac{\log(N/n)}{n}$	$R > \frac{2\log(N/n)}{\sqrt{np_0} \log\left(\frac{\log(N/n)}{np_0}\right)}$	$R > N/n^{3/2}$	$R > \frac{2\log(N/n)}{\sqrt{np_0} \log\left(\frac{\log(N/n)}{np_0}\right)}$	$R > N^{3/4}/n$
	SCAN TEST	TOT. DEG. TEST	SCAN TEST	DEG. VAR. TEST

Table 2: Polynomial time algorithms

$p_0$ known		$p_0$ unknown	
$n \lll \sqrt{N}$	$n \ggg \sqrt{N}$	$n \lll \sqrt{N}$	$n \ggg \sqrt{N}$
$R > 2\sqrt{N \log N}$	$R > N/n^{3/2}$	$R > 2\sqrt{N \log N}$	$R > N^{3/4}/n$
RELAX. SCAN TEST	TOT. DEG. TEST	RELAX. SCAN TEST	DEG. VAR. TEST

## 1.4 Finding a clique

We start the paper by addressing the problem of detecting the presence of a large clique in the graph, and treat it separately, as it is an interesting case in its own right. It is simpler and allows us to focus on the regime where  $n/\log N \rightarrow \infty$  in the rest of the paper. We establish a lower bound and prove that the following (obvious) test achieves that bound:

- *Clique number test.* This tests rejects when the size of the clique number of the graph is unusually large. It can be calibrated without knowledge of  $p_0$ , for example by permutation, but we do not know of a polynomial-time algorithm that comes even close.

## 1.5 Content

In Section 2, we consider the problem of detecting the presence of a large clique and analyze the clique number test. In Section 3, we consider the more general problem of detecting a densely connected subgraph and analyze the total degree test and the scan test. The more realistic situation of unknown  $p_0$  is handled in Section 4. In Section 5.2, we investigate polynomial-time tests. We then discuss our results and the outlook in Section 6. The technical proofs are postponed to Section 7.

## 1.6 General assumptions and notation

We assume throughout that  $N \rightarrow \infty$  and the other parameters  $n, p_0, p_1$  (and more) are allowed to change with  $N$ , unless specified otherwise. This dependency is left implicit. In particular, we assume that  $n/N \rightarrow 0$ , emphasizing the situation where the community to be detected is small compared to the size of the whole network. (When  $n$  is of the same order as  $N$ , the total degree

test is basically optimal.) We assume that  $p_0$  is bounded away from 1, which is the most interesting case by far, and that  $N^2 p_0 \rightarrow \infty$ , the latter implying that the number of edges in the network (under the null) is not bounded. We also hypothesize that either  $p_1 = 1$  or  $n \rightarrow \infty$  with  $n^2 p_1 \rightarrow \infty$ , there is a non-vanishing chance that the community does not contain any edges, precluding any test to be powerful.

We use standard notation such as  $a_n \sim b_n$  when  $a_n/b_n \rightarrow 1$ ;  $a_n = o(b_n)$  when  $a_n/b_n \rightarrow 0$ ;  $a_n = O(b_n)$  when  $a_n/b_n$  is bounded;  $a_n \asymp b_n$  when  $a_n = O(b_n)$  and  $b_n = O(a_n)$ ;  $a_n \prec b_n$  when there exists a positive constant  $C$  such that  $a_n \leq C b_n$  and  $a_n \succ b_n$  when there exists a positive constant  $C$  such that  $a_n \geq C b_n$ . For an integer  $n$  let  $n^{(2)} = n(n-1)/2$ . For two distributions  $L_1$  and  $L_2$  on the real line, let  $L_1 * L_2$  denote their convolution, which is the distribution of the sum two independent random variables  $X_1 \sim L_1$  and  $X_2 \sim L_2$ .

Because of its importance in describing the tails of the binomial distribution, the following function — which is the relative entropy or Kullback-Leibler divergence of  $\text{Bern}(q)$  to  $\text{Bern}(p)$  — will appear in our results:

$$H_p(q) = q \log \left( \frac{q}{p} \right) + (1-q) \log \left( \frac{1-q}{1-p} \right), \quad p, q \in (0, 1). \quad (2)$$

## 2 Detecting a large clique in a random graph

We start with specializing the setting to that of detecting a large clique, meaning we consider the special case where  $p_1 = 1$ . In this section,  $n$  is not necessarily increasing with  $N$ .

### 2.1 Lower bound

We establish the detection boundary, giving sufficient conditions for the problem to be too hard for any test, meaning that all tests are asymptotically powerless.

**Theorem 1.** *All tests are asymptotically powerless if*

$$\binom{N}{n} p_0^{\frac{n(n-1)}{2}} \rightarrow \infty. \quad (3)$$

The result is, in fact, very intuitive. Condition (3) implies that, with high probability under the null, the clique number is at least  $n$ , which is the size of the implanted clique under the alternative. This is a classical result in random graph theory, and finer results are known — see (Bollobás, 2001, Chap. 11). The arguments underlying Theorem 1 are, however, based on studying the likelihood ratio test when a uniform prior is assumed on the implanted clique  $S$ , which is the standard approach in detection settings; see (Lehmann and Romano, 2005, Ch. 8). In this specific setting, the second moment method — which consists in showing that the variance of the likelihood ratio tends to 0 — suffices.

### 2.2 The clique number test

Computational considerations aside, the most natural test for detecting the presence of a clique is the clique number test defined in the Introduction. We obtain the following.

**Proposition 1.** *The clique number test is powerful if*

$$\binom{N}{n} p_0^{\frac{n(n-1)}{2}} \rightarrow 0. \quad (4)$$

The proof is entirely based on the fact that, when (4) holds, the clique number under the null is at most  $n - 1$  with high probability (Bollobás, 2001, Th. 11.6), while it is at least  $n$  under the alternative. (Thus the proof is omitted.) We conclude that the clique number test is seen to achieve the detection boundary established in Theorem 1.

### 3 Detecting a dense subgraph in a random graph

We now consider the more general setting of detecting a dense subgraph in a random graph. We start with an information bound that applies to all tests, regardless of their computational requirements. We then study the total degree test and the scan test, showing that the test that combines them with a simple Bonferroni correction is essentially optimal.

#### 3.1 Lower bound

When assuming infinite computational power, what is left is the purely statistical challenge of detecting the subgraph. For simplicity, we assume that  $n$  is not too small, specifically,

$$\frac{n}{\log N} \rightarrow \infty, \quad (5)$$

though our result below partially extends to this, particularly when  $p_1$  is constant. As usual, a minimax lower bound is derived by choosing a prior over the composite alternative. Assuming that  $p_0$  and  $p_1$  are known, because of symmetry, the uniform prior over the community  $S$  is least favorable, so that we consider testing

$$H_0 : \mathcal{G} \sim \mathbb{G}(N, p_0) \text{ versus } \bar{H}_1 : \mathcal{G} \sim \mathbb{G}(N, p_0; n, p_1), \quad (6)$$

where the latter is the model where the community  $S$  is chosen uniformly at random among subset of nodes of size  $n$ , and then for  $i \neq j$ ,  $\mathbb{P}(W_{i,j} = 1) = p_1$  if  $i, j \in S$ , while  $\mathbb{P}(W_{i,j} = 1) = p_0$  otherwise. For this simple versus simple testing problem, the likelihood ratio test is optimal, which is what we examine to derive the following lower bound. Remember the entropy function defined in (2).

**Theorem 2.** *Assuming (5) and (1) hold, all tests are asymptotically powerless if*

$$\frac{p_1 - p_0}{\sqrt{p_0}} \frac{n^2}{N} \rightarrow 0, \quad (7)$$

and

$$\limsup \frac{nH(p_1)}{2 \log(N/n)} < 1. \quad (8)$$

Conditions (7) and (8) have their equivalent in the work of Butucea and Ingster (2011). That said, (8) is more complex here because of the different behaviors of the entropy function according to whether  $p_1/p_0$  is small or large — corresponding to the difference between large deviations and moderate deviations of the binomial distribution. Only in the case where  $p_1/p_0 \rightarrow 1$  is the normal approximation to the binomial in effect.

To better appreciate (8), note that it is equivalent to

$$\limsup \frac{(p_1 - p_0)^2}{4p_0(1 - p_0)} \frac{n}{\log(N/n)} < 1, \quad \text{when } \frac{np_0}{\log(N/n)} \rightarrow \infty; \quad (9)$$

and

$$\limsup \frac{p_1}{2(1-p_0)} \frac{n}{\log(N/n)} \log \left( \frac{\log(N/n)}{np_0} \right) < 1, \quad \text{when } \frac{np_0}{\log(N/n)} \rightarrow 0. \quad (10)$$

In (9),  $np_0$  is larger and only the moderate deviations of the binomial distribution are involved, while in (10),  $np_0$  is smaller and the large deviations come into play.

Theorem 2 happens to be sharp because, as we show next, the test that combines the total degree test and the scan test is asymptotically powerful when the conditions (7) and (8) are — roughly speaking — reversed.

### 3.2 The total degree test

The total degree test rejects for large values of

$$W := \sum_{1 \leq i < j \leq N} W_{i,j}. \quad (11)$$

The resulting test is exceedingly simple to analyze, since

$$W \sim \text{Bin}(N^{(2)} - n^{(2)}, p_0) * \text{Bin}(n^{(2)}, p_1). \quad (12)$$

**Proposition 2.** *The total degree tests is powerful if*

$$\frac{p_1 - p_0}{\sqrt{p_0}} \frac{n^2}{N} \rightarrow \infty. \quad (13)$$

It is equally straightforward to show that the total degree has risk strictly less than one — meaning has some non-negligible power — when the same ratio tends to a positive and finite constant, while it is asymptotically powerless when that ratio tends to zero.

### 3.3 The scan test

The scan test is another name for the generalized likelihood ratio test, and corresponds to the test that is based on the maximum modularity. It is particularly simple when  $p_0$  is known, as it rejects for large values of

$$W_{[n]}^* := \max_{|S|=n} W_S, \quad W_S := \sum_{i,j \in S, i < j} W_{i,j}. \quad (14)$$

Unlike the total degree (11), the scan statistic (14) has an intricate distribution as the partial sums  $W_S$  are not independent. Nevertheless, the union bound and standard tail bounds for the binomial distribution lead to the following result.

**Proposition 3.** *The scan test is powerful if*

$$\liminf \frac{nH(p_1)}{2\log(N/n)} > 1. \quad (15)$$

### 3.4 The combined test

Having studied these two tests individually, we are now in a position to consider them together, by which we mean a simple Bonferroni combination which rejects when either of the two tests rejects. Looking back at our lower bound and the performance bounds we established for these tests, we come to the following conclusion. When the limit in (7) is infinite — yielding (13) — then the total degree test is asymptotically powerful by Proposition 2. When the limit inferior in (8) exceeds one — yielding (15) — then the scan test is asymptotically powerful by Proposition 3.



### 3.5 Adaptation to unknown $n$

The scan statistic in (14) requires knowledge of  $n$ . When this is unknown, the common procedure is to combine the scan tests at all different sizes  $n$  using a simple Bonferroni correction. This is done in (Butucea and Ingster, 2011), with the conclusion that the resulting test is essentially as powerful as the individual tests. It is straightforward to see that, here too, the tail bound used in the proof of Proposition 3 allows for enough room to scan over all subgraphs of all sizes.

## 4 When $p_0$ is unknown: the fixed expected total degree model

Although it leads to interesting mathematics, the setting where  $p_0$  is known is, for the most part, impractical. In this section, we evaluate how not knowing  $p_0$  changes the difficulty of the problem. In fact, it makes the problem strictly more difficult in the denser regime.

There are (at least) two ways of formalizing the situation where  $p_0$  is unknown. In the first option, we still consider the exact same hypothesis testing problem, but maximize the risk over relevant subsets of  $p_0$ 's and  $p_1$ 's, since now even the null hypothesis is composite. In the second option — which is the one we detail — for a given pair of probabilities  $0 < p'_0 \leq p_1 < 1$ , we consider testing

$$H_0 : \mathcal{G} \sim \mathbb{G}(N, p_0) \text{ versus } \bar{H}'_1 : \mathcal{G} \sim \mathbb{G}(N, p'_0; n, p_1), \quad p_0 := p'_0 + (p_1 - p'_0) \frac{n^{(2)}}{N^{(2)}}. \quad (16)$$

Note that, in this setting, we still assume that  $p_0, p_1, n$  are known to the statistician. By design, the graph has the same expected total degree under the null and under the alternative hypotheses, that is we have

$$\mathbb{E}_0(W) = N^{(2)}p_0 + n^{(2)}(1 - p_0) = \mathbb{E}'_S(W), \quad \forall S : |S| = n,$$

where  $\mathbb{P}'_S$  and  $\mathbb{E}'_S$  denote the probability distribution and corresponding expectation under the model where, for any  $i \neq j$ ,  $\mathbb{P}(W_{i,j} = 1) = p_1$  if  $i, j \in S$ , while  $\mathbb{P}(W_{i,j} = 1) = p'_0$  otherwise.

The risk of a test  $T$  for this problem is defined as

$$\gamma'_N(T) = \mathbb{P}_0(T = 1) + \max_{|S|=n} \mathbb{P}'_S(T = 0).$$

We say that the a sequence of tests  $(T_N)$  is asymptotically powerful for the problem with fixed expected total degree (resp. powerless) if  $\gamma'_N(T_N) \rightarrow 0$  (resp.  $\gamma'_N(T_N) \rightarrow 1$ ).

We first compute the detection boundary for this problem and then exhibit some tests achieving this detection boundary. Interestingly, these tests do not require the knowledge of  $p_0$  and  $p_1$ , or even  $n$ , so that they can be used in the original setting (6) when these parameters are unknown.

### 4.1 Lower bound

**Theorem 3.** *Assuming (5) holds and that*

$$\log \left( 1 \vee \frac{1}{np'_0} \right) = o \left[ \log \left( \frac{N}{n} \right) \right], \quad (17)$$

*all tests are asymptotically powerless for the problem (16) if*

$$\frac{p_1 - p'_0}{\sqrt{p'_0}} \frac{n^{3/2}}{N^{3/4}} \rightarrow 0 \quad (18)$$



and

$$\limsup \frac{nH_{p'_0}(p_1)}{2\log(N/n)} < 1. \quad (19)$$

Comparing with Theorem 2, where  $p_0$  is assumed to be known, the condition (18) is substantially weaker than the corresponding condition (7), while we shall see in the proof that (19) is comparable to (8). That said, when  $n^2 < N$ , the entropy condition (8) is a stronger requirement than either (7) or (18), implying that the setting where  $p_0$  is known and the setting where unknown are asymptotically as difficult in that case.

## 4.2 Degree variance test

By construction, the total degree  $W$  has the same expectation under the null and under the alternative in the testing problem with fixed expected total degree — and same variance also up to second order — making it difficult to see how to fruitfully use this statistic in this context.

We design instead a test based on comparing the two estimators for the node degree variance, not unlike an analysis of variance. Let

$$W_{i\cdot} = \sum_{j \neq i} W_{i,j} \quad (20)$$

denote the degree of node  $i$  in the whole network. The first estimate is simply the maximum likelihood estimator under the null

$$V_1 = (N-1) \frac{N^{(2)}}{N^{(2)}-1} \hat{p}_0(1-\hat{p}_0), \quad \hat{p}_0 := \frac{W}{N^{(2)}}.$$

The second estimator is some sort of sample variance, modified to account for the fact that the  $W_{i\cdot}$  are not independent

$$V_2 = \frac{1}{N-2} \sum_{i=1}^N (W_{i\cdot} - (N-1)\hat{p}_0)^2.$$

Both estimators are unbiased for the degree variance under the null, meaning,  $\mathbb{E}_0 V_1 = \mathbb{E}_0 V_2 = (N-1)p_0(1-p_0)$ . Under the alternative,  $V_2$  tends to be larger than  $V_1$ , leading to a test that rejects for large values of

$$V^* := \frac{V}{\sqrt{N\hat{p}_0}}, \quad V := V_2 - V_1. \quad (21)$$

**Proposition 4.** *Assume that  $p_0 \succ 1/N$ . The degree variance test is asymptotically powerful under fixed expected total degree if*

$$\frac{(p_1 - p'_0)^2}{p'_0} \frac{n^3}{N^{3/2}} \rightarrow \infty \quad (22)$$

The test based on  $V^*$  achieves the part (18) of the detection boundary. We note that computing  $V^*$  does not require knowledge of  $p_0$ ,  $p_1$  or  $n$ , and in fact, its calibration can be done without any knowledge of these parameters via a form of parametric bootstrap, as we do for the scan test below.

## 4.3 The scan test

When  $p_0$  is not available a priori, we have at least three options:

- *Estimate  $p_0$ .* We replace  $p_0$  with its maximum likelihood estimator under the null, i.e.,  $\hat{p}_0 = W/N^{(2)}$ , and then compare the magnitude of the observed scan statistic (14) with what one would get under a random graph model with probability of connection equal to  $\hat{p}_0$ .
- *Generalized likelihood ratio test.* We simply implement the actual generalized likelihood ratio test (Kulldorff, 1997), which rejects for large values of

$$\max_{|S|=n} \left[ n^{(2)} h(\hat{p}_{1,S}) + (N^{(2)} - n^{(2)}) h(\hat{p}_{0,S}) - N^{(2)} h(\hat{p}_0) \right] ,$$

where  $h(p) := p \log p + (1 - p) \log(1 - p)$ ,  $\hat{p}_0$  as above, and

$$\hat{p}_{1,S} := \frac{W_S}{n^{(2)}}, \quad \hat{p}_{0,S} := \frac{W - W_S}{N^{(2)} - n^{(2)}} ,$$

which are the maximum likelihood estimates of  $p_1$  and  $p_0$  for a given subset  $S$ .

- *Calibration by permutation.* We compare the observed value of the scan statistic to simulated values obtained by generating a random graph with either the same number of edges — which leads to a calibration very similar to the first option — or the same degree distribution — which is the basis for in the modularity function of Newman and Girvan (2004).

We focus on the first option.

**Proposition 5.** *Assume that  $\liminf p_0 N^2/n > 1$ . The scan test calibrated by estimation of  $p_0$  is asymptotically powerful for fixed expected total degree if*

$$\liminf \frac{nH(p_1)}{2 \log(N/n)} > 1 . \quad (23)$$

Hence, the scan test calibrated by estimation of  $p_0$  achieves the entropy condition (8) without requiring the knowledge of  $p_0$  or  $p_1$ . We mention that adaptation to unknown  $n$  may be achieved as described in Section 3.5.

#### 4.4 Combined test and full adaptation to unknown $p_0$

A combination of the degree variance test and of the scan test calibrated by estimation of  $p_0$  is seen to achieve the detection boundary established in Theorem 3, without requiring knowledge of  $p_0$  or  $p_1$ , or even  $n$ .

### 5 Testing in polynomial-time

While computing the total degree (11) or the degree variance statistic (21) can be done in linear time in the size of the network, i.e., in  $O(N^2)$  time, computing the scan statistic (14) is combinatorial in nature and there is no known polynomial-time algorithm to compute it. To see this, note that the ability to compute (14) in polynomial-time implies the ability to compute the size of the largest clique in the graph, since this is equal to

$$\max\{n : W_{[n]}^* = n^{(2)}\} ,$$

and computing the size of the largest clique in a general graph is known to be NP-hard (Karp, 1972), and even hard to approximate (Zuckerman, 2006).

A question of particular importance in modern times is determining the tradeoff between statistical performance and computational complexity. At the most basic level, this boils down to answering the following question: *What can be done in polynomial-time?*

## 5.1 Convex relaxation scan test

We now suggest a convex relaxation to the problem of computing the scan statistic. To do so, we follow the footsteps of [Berthet and Rigollet \(2012\)](#), who consider the problem of detecting a sparse principal component based on a sample from a multivariate Gaussian distribution in dimension  $N$ . Assuming the sparse component has at most  $n$  nonzero entries, they show that a near-optimal procedure is based on the largest eigenvalue of any  $n$ -by- $n$  submatrix of the sample covariance matrix. Computing this statistic is NP-hard, so they resort to the convex relaxation of [d'Aspremont et al. \(2007\)](#), which they also study. We apply their procedure to  $\mathbf{W}^2$ .

Formally, for a positive semidefinite matrix  $\mathbf{B} \in \mathbb{R}^{N \times N}$  and  $1 \leq n \leq N$ , define

$$\lambda_n^{\max}(\mathbf{B}) = \max_{|S|=n} \lambda^{\max}(\mathbf{B}_S) ,$$

where  $\mathbf{B}_S$  denotes the principal submatrix of  $\mathbf{B}$  indexed by  $S \subset \{1, \dots, N\}$  and  $\lambda^{\max}(\mathbf{B})$  the largest eigenvalue of  $\mathbf{B}$ . [d'Aspremont et al. \(2007\)](#) relaxed this to

$$\text{SDP}_n(\mathbf{B}) = \max_{\mathbf{Z}} \text{Trace}(\mathbf{B}\mathbf{Z}), \quad \text{subject to } \mathbf{Z} \succeq 0, \text{Trace}(\mathbf{Z}) = 1, |\mathbf{Z}|_1 \leq n ,$$

where the maximum is over positive semidefinite matrices  $\mathbf{Z} = (Z_{st}) \in \mathbb{R}^{N \times N}$  and  $|\mathbf{Z}|_1 = \sum_{s,t} |Z_{st}|$ . We consider the relaxed scan test, which rejects for large values of

$$\text{SDP}_n(\mathbf{W}^2) . \tag{24}$$

When  $p_0$  is known, we simply calibrate the procedure by Monte Carlo simulations, effectively generating  $\mathbf{W}_1, \dots, \mathbf{W}_B$  i.i.d. from  $\mathbb{G}(N, p_0)$  and computing  $\text{SDP}_n(\mathbf{W}_b^2)$  for each  $b = 1, \dots, B$ , and estimating the p-value by the fraction of  $b$ 's such that  $\text{SDP}_n(\mathbf{W}_b^2) \geq \text{SDP}_n(\mathbf{W}^2)$ . Typically  $B$  is a large number, and below we consider the asymptote where  $B = \infty$ .

When  $p_0$  is unknown, we estimate  $p_0$  as we did for the scan test in [Proposition 5](#), and then calibrate the statistic by Monte Carlo, effectively using a form of parametric bootstrap.

In either case, we have the following.

**Proposition 6.** *Assume that (1) holds and  $n \leq N^{1/2-t}$  for some  $t > 0$ . Then, the relaxed scan test is powerful if*

$$\liminf \frac{n}{\sqrt{N \log(N)}} \frac{(p_1 - p_0)^2}{p_0} > 2 . \tag{25}$$

To gain some insights on the relative performance of the scan test and the relaxed scan test, let us assume that  $n^2 \ll N$ , and  $np_0 \gg \log(N/n)$ . Applying [Proposition 3](#) (or [Proposition 5](#)) in this setting, we find that the scan test is asymptotically powerful when

$$\frac{(p_1 - p_0)^2}{p_0} \succ \frac{\log(N/n)}{n} .$$

Thus, comparing with (25), we lose a factor  $\sqrt{N/\log(N)}$  when using the relaxed version. In the denser regime where  $n^2 \gg N \log(N)$ , the total degree test and degree variance test both have stronger theoretical guarantees established in [Proposition 2](#) and [Proposition 4](#) respectively. Below we explain why the  $\sqrt{N/\log(N)}$  loss is not unexpected.

## Optimality

The problem  $H_0 : \mathcal{G} \sim \mathbb{G}(N, 1/2)$  versus  $H_1 : \mathcal{G} \sim \mathbb{G}(N, 1/2; n, 1)$  is called the Planted (or Hidden) Clique Problem (Feige and Ron, 2010) and has become one of the most emblematic statistical problems where computational constraints seem to substantially affect the difficulty of the problem. Recent advances in compressed sensing and matrix completion have shown that computationally tractable algorithms can achieve the absolute information bounds (up to constants) in most cases. In contrast, in the Planted (or Hidden) Clique Problem there is no known polynomial-time algorithm that can detect a clique of size  $n = o(\sqrt{N})$  (Dekel et al., 2011), while the clique test can detect a clique of size  $n \asymp \log N$ , as shown in Proposition 1. In fact, the problem is provably hard in some computational models, such as monotone circuits (Feldman et al., 2012; Rossman, 2010). We refer to Berthet and Rigollet (2012) for a thorough discussion.

More generally, we may want to characterize the sequences  $(n, N, p_0, p_1)$  for which there are asymptotically powerful tests running in polynomial time. In our findings, the only situation where we found this to be true was in the dense regime, where the total degree test is both powerful in the large-sample limit and computable in polynomial time. (Replace this with the degree variance test when  $p_0$  is unknown.)

## 5.2 Other polynomial-time tests

### 5.2.1 The maximum degree test

Perhaps the first computationally-feasible test that comes to mind in the sparse regime is the test based on the maximum degree

$$\max_{i=1, \dots, N} W_{i\cdot} \ , \quad (26)$$

where  $W_{i\cdot}$  is the degree of node  $i$  in the graph, defined in (20).

**Proposition 7.** *The maximal degree test is asymptotically powerful if  $p_0 \gg \log(N)/N$  and*

$$\liminf \frac{n^2}{N \log(N)} \frac{(p_1 - p_0)^2}{p_0(1 - p_0)} > 2 \ .$$

*Under condition (1), the maximal degree test is asymptotically powerless if  $\limsup \log(n)/\log(N) < 1$  and*

$$\frac{n^2}{N \log(N)} \frac{(p_1 - p_0)^2}{p_0(1 - p_0)} \rightarrow 0 \quad (27)$$

Comparing with Propositions 2 and 6, we observe that the maximum degree test is either less powerful than the relaxed scan test (when  $n \leq N^{1/2-t}$  for any  $t > 0$ ) or less powerful than the total degree test (when  $n \gg \sqrt{N/\log(N)}$ ). For unknown  $p_0$ , the maximum degree test is also less powerful than the degree variance test.

### 5.2.2 Densest subgraph test

Another possible avenue for designing computationally tractable tests for the problem at hand lies in algorithms for finding dense subgraphs of a given size. We follow (Khuller and Saha, 2009), where the reader will find appropriate references and additional results. Define the density of a subgraph  $S \subset \mathcal{V}$  as

$$h(S) = \frac{|E_S|}{|S|}, \quad \text{where } E_S = \{(i, j) \in S^2 : W_{i,j} = 1\} \ .$$

Finding  $S \subset \mathcal{V}$  that maximizes  $h(S)$  may be done in polynomial-time.

**Proposition 8.** *Assume that  $p_0 \gg \log(N)/N$ .*

1. *Under the null hypothesis,  $\max_S h(S) \sim_{\mathbb{P}_0} h(\mathcal{V}) \sim Np_0/2$  and this maximum is achieved at subsets  $S$  satisfying  $|S| \sim N$ .*
2. *The densest subgraph test is powerful if  $\liminf \frac{np_1}{Np_0} > 1$ .*
3. *Assume that  $\frac{np_1}{Np_0} \rightarrow 0$ . Under the alternative hypothesis,  $\max_S h(S) \sim_{\mathbb{P}_S} h(\mathcal{V}) \sim_{\mathbb{P}_S} Np_0/2$  and this maximum is achieved at subsets  $S$  satisfying  $|S| \sim N$ .*

The condition  $\liminf \frac{np_1}{Np_0} > 1$  is stronger than what we have obtained for the relaxed scan test (25) in the sparser case ( $n \leq N^{1/2-t}$  for any  $t > 0$ ) and than what we have obtained for the total degree test (13) and the degree variance test (22) in the less sparse case ( $n \gg \sqrt{N}$ ). If  $np_1/Np_0 \rightarrow 0$ , then the densest subgraph statistic seems to behave like the total degree statistic and we therefore expect similar performances although we have no proof of this statement.

In order to improve the power, we would like to restrict our attention to subgraphs of size  $n$  (assumed known for now) and use  $\max_{|S|=n} h(S)$ . Computing this, however, is NP-hard, and there is no known polynomial-time approximation within a constant factor. Nevertheless, the following variant statistic  $\max_{|S| \geq n} h(S)$  can be approximated within a constant factor in polynomial-time. However, the power of the resulting test is not improved. Since the statistic  $\max_{|S| \geq n} h(S)$  may only be approximated within a constant factor, the resulting test is powerful only if  $np_1 \geq CNp_0$  where  $C$  is positive constant that depends on this approximation factor.

## 6 Discussion

With this paper, we have established the fundamental statistical (information theoretic) difficulty of detecting a community in a network, modeled as the detection of an unusually dense subgraph within an Erdős-Rényi random graph, in the quasi-normal regime where  $np_0$  is not too small as made explicit in (1). The quasi-Poisson regime, where  $np_0$  is smaller, requires different arguments and the application of somewhat different tests, and this will be detailed in a separate paper under preparation.

For the time being, in the quasi-normal regime, we learned the following. In the moderately sparse setting —  $n \gg N^{2/3}$  for known  $p_0$  and  $n \gg N^{3/4}$  for unknown  $p_0$  — this detection boundary is achieved by polynomial-time tests. In the sparser setting, there is a large discrepancy between the information theoretic boundaries and performances of known polynomial tests, which in view of the Planted Clique Problem, is not surprising.

It is of great interest to study this optimal detection boundary, this time under computational constraints, a theme of contemporary importance in statistics, machine learning and computer science. This promisingly rich line of research is well beyond the scope of the present paper.

## 7 Proofs

### 7.1 Auxiliary results

The following is Chernoff's bound for the binomial distribution. Remember the definition of  $H$  in (2).

**Lemma 1** (Chernoff's bound). *For any positive integer  $n$ , any  $q, p_0 \in (0, 1)$ , we have*

$$\mathbb{P}(\text{Bin}(n, p_0) \geq qn) \leq \exp(-nH(q)). \quad (28)$$

A consequence of Chernoff's bound is Bernstein's inequality for the binomial distribution.

**Lemma 2** (Bernstein's inequality). *For positive integer  $n$ , any  $p_0 \in (0, 1)$  and any  $x > 0$ , we have*

$$\mathbb{P}[\text{Bin}(n, p_0) \geq np_0 + x] \leq \exp\left[-\frac{x^2}{2[np_0(1-p_0) + x/3]}\right].$$

We will need the following basic properties of the entropy function.

**Lemma 3.** *For  $p_0 \in (0, 1)$ ,  $H(q)$  is convex in  $q \in [0, 1]$ . Moreover,*

$$H_p(q) = \begin{cases} \frac{(q-p)^2}{2p(1-p)} + O\left(\frac{(q-p)^3}{p^2}\right), & \frac{q}{p} \rightarrow 1; \\ p(r \log r - r + 1), & \frac{q}{p} \rightarrow r \in (1, \infty), \quad p \rightarrow 0; \\ q \log\left(\frac{q}{p}\right) + O(q), & \frac{q}{p} \rightarrow \infty. \end{cases} \quad (29)$$

We will also use the following upper bound on the binomial coefficients.

**Lemma 4.** *For any integers  $1 \leq k \leq n$ ,*

$$k \log(n/k) \leq \log \binom{n}{k} \leq k \log(ne/k), \quad (30)$$

where  $e = \exp(1)$ .

The next result bounds the hypergeometric distribution with the corresponding binomial distribution. Let  $\text{Hyp}(N, m, n)$  denotes the hypergeometric distribution counting the number of red balls in  $n$  draws from an urn containing  $m$  red balls out of  $N$ .

**Lemma 5.**  *$\text{Hyp}(N, m, n)$  is stochastically smaller than  $\text{Bin}(n, m/(N-m))$ .*

*Proof.* Suppose the balls are picked one by one without replacement. At each stage, the probability of selecting a red ball is smaller than  $m/(N-m)$ . The result follows.  $\square$

## 7.2 Proof of Theorem 1

Following standard lines, we start by reducing the composite alternative to a simple alternative by considering the uniform prior  $\pi$  on subsets  $S \subset [N] := \{1, \dots, N\}$  of size  $|S| = n$ . The resulting likelihood ratio is

$$L = \frac{\#\{S \subset [N] : |S| = n, W_S = n^{(2)}\}}{\binom{N}{n} p_0^{n(n-1)/2}}, \quad (31)$$

which is the observed number of cliques of size  $n$  divided by the expected number under the null.

The risk of any test for the original problem is well-known to be bounded from below by the risk of the likelihood ratio test  $\{L > 1\}$  for this 'averaged' problem, which is equal to

$$\gamma_L := \mathbb{P}_0(L > 1) + \mathbb{E}_0(L\{L \leq 1\}).$$

Therefore, it suffices to show that  $\gamma_L \rightarrow 1$ . Here we use arguably the simplest method, a second moment argument, which is based on the fact that

$$\gamma_L = 1 - \mathbb{E}_0 |L - 1| \geq 1 - \sqrt{\text{Var}_0(L)},$$

by the Cauchy-Schwarz inequality, so that it is enough to prove that  $\text{Var}_0(L) \rightarrow 0$ . We do so by showing that  $\mathbb{E}_0(L^2) \leq 1 + o(1)$ .

Note that

$$L = p_0^{-n^{(2)}} \pi \left[ W_S = n^{(2)} \right],$$

where  $\pi[\cdot]$  denotes the expectation with respect to  $\pi$ . Hence, by Fubini's theorem, we have

$$\mathbb{E}_0 L^2 = \pi^{\otimes 2} \left[ p_0^{-2n^{(2)}} \mathbb{P}_0(W_{S_1} = W_{S_2} = n^{(2)}) \right] = \pi^{\otimes 2} \left[ p_0^{-K(K-1)/2} \right],$$

where  $K := |S_1 \cap S_2|$ . Indeed, the event  $\{W_{S_1} = W_{S_2} = n^{(2)}\}$  means that all edges between pairs of nodes in  $S_1$  exist, and similarly for  $S_2$ , and there are a total of  $n(n-1) + K(K-1)/2$  such edges.

Before going further, note that (3) and (30) imply that

$$\log(N/n) - \frac{(n-1)}{2} \log(1/p_0) \rightarrow \infty. \quad (32)$$

In particular, this means that  $n \leq 3 \log N$ , eventually, and therefore

$$\frac{n^2}{N} = O((\log N)^2/N) \rightarrow 0. \quad (33)$$

Since  $K \sim \text{Hyp}(N, n, n)$ , by Lemma 5,  $K$  is stochastically bounded by  $\text{Bin}(n, \rho)$ , where  $\rho := n/(N-n)$ . Hence, with and Lemma 1, we have

$$\begin{aligned} \mathbb{P}(K \geq k) &\leq \mathbb{P}(\text{Hyp}(N, n, n) \geq k) \\ &\leq \mathbb{P}(\text{Bin}(n, \rho) \geq k) \\ &\leq \exp(-nH_\rho(k/n)). \end{aligned} \quad (34)$$

Now, using Lemma 3 and (33), for  $k \geq 2$  we get

$$nH_\rho(k/n) = k \log(k/(n\rho)) + O(k) = k \log(kN/n^2) + O(k).$$

Hence,

$$\begin{aligned} \pi^{\otimes 2} \left[ p_0^{-K(K-1)/2} \right] &= \mathbb{P}_0(K \leq 1) + \sum_{k=2}^n \exp \left( \frac{k(k-1)}{2} \log(1/p_0) - nH_\rho(k/n) \right) \\ &\leq 1 + \sum_{k=2}^n \exp \left( k \left[ \frac{(k-1)}{2} \log(1/p_0) - \log(kN/n^2) + O(1) \right] \right). \end{aligned} \quad (35)$$

For  $a > 0$  fixed, the function  $x \rightarrow ax - \log x$  is decreasing on  $(0, 1/a)$  and increasing on  $(1/a, \infty)$ . Therefore,

$$\frac{(k-1)}{2} \log(1/p_0) - \log(kN/n^2) \leq -\omega,$$

where

$$\omega := \min \left( \log(N/n^2) - \frac{1}{2} \log(1/p_0), \log(N/n) - \frac{n-1}{2} \log(1/p_0) \right).$$

By (32), the second term in the maximum tends to  $\infty$ . This also the case of the first term, since

$$\log(N/n^2) - \frac{1}{2} \log(1/p_0) = \log(N/n) - \frac{n-1}{2} \log(1/p_0) + \frac{n}{2} \log(1/p_0) - \log n,$$



with the second difference bounded from below. Hence,  $\omega \rightarrow \infty$ . Hence, the sum in (35) is bounded by

$$\sum_{k=2}^n \exp(k[\omega + O(1)]) \leq \frac{e^{-\omega/2}}{1 - e^{-\omega/2}} \rightarrow 0,$$

eventually.

Hence we showed that  $\mathbb{E}_0(L^2) \leq 1 + o(1)$  and the proof of Theorem 1 is complete.

### 7.3 Proof of Theorem 2

We assume that (1), (7) and (8) hold. We reduce the composite alternative to a simple alternative by considering the uniform prior  $\pi$  on subsets  $S \subset [N] := \{1, \dots, N\}$  of size  $|S| = n$ . The resulting likelihood ratio is

$$L(A) = \binom{N}{n}^{-1} \sum_{|S|=n} L_S(A) = \pi[L_S(A)], \quad (36)$$

where  $\pi[\cdot]$  is the expectation with respect to  $S \sim \pi$ ,  $A = (W_{i,j} : 1 \leq i < j \leq N)$  and

$$L_S := \exp(\theta W_S - \Lambda(\theta)n^{(2)}), \quad (37)$$

with

$$\theta := \theta_{p_1}, \quad \theta_q := \log \left( \frac{q(1-p_0)}{p_0(1-q)} \right) \quad (38)$$

and

$$\Lambda(\theta) := \log(1 - p_0 + p_0 e^\theta),$$

which is the moment generating function of  $\text{Bern}(p_0)$ .

Still leaving  $p_0$  implicit, let  $H_{p_0}(q)$  be short for  $H(q)$ . It is well-known that  $H$  is the Fenchel-Legendre transform of  $\Lambda$ ; more specifically, for  $q \in (p_0, 1)$ ,

$$H(q) = \sup_{\theta \geq 0} [q\theta - \Lambda(\theta)] = q\theta_q - \Lambda(\theta_q). \quad (39)$$

The second moment argument used in Section 7.2 is also applicable here, though it does not yield sharp bounds. In Case 1 below (see Subsection 7.3.3), which is the regime where the moderate deviations of the binomial come into play, this method leads to a requirement that the limit superior in (8) be bounded by  $1/2$  instead of  $1$ . And, worse than that, in Case 3 below, which is the regime where the large deviations of the binomial are involved, it does not provide any useful bound whatsoever.

Fortunately, a finer approach was suggested by Ingster (1997). The refinement is based on bounding the first and second moments of a truncated likelihood ratio. Here we follow Butucea and Ingster (2011). They work with the following truncated likelihood

$$\tilde{L} = \binom{N}{n}^{-1} \sum_{|S|=n} \mathbb{1}_{\Gamma_S} L_S.$$

where the events  $\Gamma_S$  will be specified below. We note  $\Gamma = \bigcap_{|S|=n} \Gamma_S$ . Using the triangle inequality, the fact that  $\tilde{L} \leq L$  and the Cauchy-Schwarz inequality, we have the following upper bound:

$$\begin{aligned} \mathbb{E}_0 |L - 1| &\leq \mathbb{E}_0 |\tilde{L} - 1| + \mathbb{E}_0 (L - \tilde{L}) \\ &\leq \sqrt{\mathbb{E}_0 [\tilde{L}^2] - 1} + 2(1 - \mathbb{E}_0 [\tilde{L}]) + (1 - \mathbb{E}_0 [\tilde{L}]), \end{aligned}$$

so that  $\gamma_L \rightarrow 1$  when  $\mathbb{E}_0[\tilde{L}^2] \rightarrow 1$  and  $\mathbb{E}_0[\tilde{L}] \rightarrow 1$ . Note that contrary to Butucea and Ingster (2011), we do not require that  $\mathbb{P}_0(\Gamma) \rightarrow 1$ . More precisely, we shall prove that  $(1, 1)$  is an accumulation point of any subsequence of  $(\mathbb{E}_0 \tilde{L}, \mathbb{E}_0[\tilde{L}^2])$ . Adopting this approach allows us to assume that  $p_1/p_0$  converges to  $r \in [1, \infty]$ ,  $p_1^2/p_0$  converges to  $r_2 \in [0, \infty]$  and that

$$\frac{nH(p_1)}{2\log(N/n)} < 1 - \eta_0, \quad (40)$$

for some  $\eta_0 \in (0, 1)$  fixed. Notice that (5) and (8) imply that  $H(p_1) \rightarrow 0$ , which by Lemma 3 forces either  $p_1/p_0 \rightarrow 1$  or  $p_1 \rightarrow 0$ ; in any case,  $p_1$  is bounded away from 1 this time.

In what follows, we provide the general arguments while the proof of the technical results (Lemmas 6-8) is postponed to the end of the section. To show these technical results, we divide the analysis depending on the behaviour of  $p_1/p_0$

$$\frac{p_1}{p_0} \rightarrow \begin{cases} r = 1, & (41) \\ r \in (1, \infty), & (42) \\ r = \infty. & (43) \end{cases}$$

In regime (41), the moderate deviations of the binomial distribution dominate and these are asymptotically equivalent to normal (Gaussian) deviations; in particular, it is in this setting (with  $p_0$  constant) that Butucea and Ingster (2011) successfully reduce the binary setting to the normal setting. In regime (43), the large deviations of the binomial distribution dominate, which are not alike the normal deviations and lead to a completely different regime. Regime (42) is intermediary and requires special treatment.

First, we need some notations to introduce  $\Gamma_S$ . Define the numbers

$$k_* = \left\lceil 1 + 2 \frac{\log(N/n)}{\log\left(1 + \frac{(p_1-p_0)^2}{p_0(1-p_0)}\right)} \right\rceil \wedge n, \quad (44)$$

$$k_{\min} = \left\lceil 1 + 2 \frac{\log\left(\frac{Nk_*}{n^2}\right) - \log\left\{\log\left(\frac{n}{\log(N/n)}\right) \wedge \log(N/n)\right\}}{\log\left(1 + \frac{(p_1-p_0)^2}{p_0(1-p_0)}\right)} \right\rceil \wedge n. \quad (45)$$

The exact expression of  $k_{\min}$  will be useful for bounding the second moment of  $\tilde{L}$ . For the time being, we only need to have in mind the properties summarized in the following lemma.

**Lemma 6.** *We have  $k_{\min} \rightarrow \infty$ ,  $k_{\min} \sim k_*$ , and  $\log(n/k_{\min}) = o[\log(N/n)]$ .*

We define  $\Gamma_S$  as follows

$$\Gamma_S := \bigcap_{k=\lfloor k_{\min} \rfloor + 1}^n \{W_T \leq w_k, \forall T \subset S \text{ such that } |T| = k\}, \quad (46)$$

where  $w_k := q_k k^{(2)}$ , with

$$\frac{(k-1)}{2} H(q_k) = \log(N/k) + 2. \quad (47)$$

This construction is possible by the following lemma, which serves as a definition.

**Lemma 7.** *For any integer  $k$  between  $k_{\min} + 1$  and  $n$ , there exists a unique  $q_k \in (p_0, 1)$  such that*

$$\frac{(k-1)}{2} H(q_k) = \log(N/k) + 2.$$

Moreover,  $q_k$  satisfies  $\theta_{q_k} \leq 2\theta$ .

### 7.3.1 First truncated moment

We first prove that  $\mathbb{E}_0 \tilde{L} \rightarrow 1$ . By Fubini's theorem, we have

$$\mathbb{E}_0 \tilde{L} = \pi[\mathbb{E}_0[L_S \mathbb{1}_{\Gamma_S}]] = \pi[\mathbb{P}_S(\Gamma_S)] = \mathbb{P}_S(\Gamma_S),$$

where  $S$  is any fixed subset of size  $n$  in  $\{1, \dots, N\}$  and this last inequality is by the fact that  $\mathbb{P}_S(\Gamma_S)$  does not depend on  $S$  by symmetry. By the union bound, Chernoff's bound (28) and (30),

$$\begin{aligned} 1 - \mathbb{P}_S(\Gamma_S) &\leq \sum_{k=\lfloor k_{\min} \rfloor + 1}^n \sum_{T \subset S, |T|=k} \mathbb{P}_S(W_T > q_k k^{(2)}) \\ &\leq \sum_{k=\lfloor k_{\min} \rfloor + 1}^n \binom{n}{k} \mathbb{P}(\text{Bin}(k^{(2)}, p_1) > q_k k^{(2)}) \\ &\leq \sum_{k=\lfloor k_{\min} \rfloor + 1}^n \exp \left[ k \left( \log(ne/k) - \frac{(k-1)}{2} H_{p_1}(q_k) \right) \right]. \end{aligned}$$

We then conclude that  $1 - \mathbb{P}_S(\Gamma_S) = o(1)$  using the following result.

**Lemma 8.** *We have*

$$\min_{k=\lfloor k_{\min} \rfloor + 1, \dots, n} \left( \frac{k-1}{2} H_{p_1}(q_k) - \log \left( \frac{n}{k} \right) \right) \rightarrow \infty. \quad (48)$$

### 7.3.2 Second truncated moment

We now prove that  $\mathbb{E}_0 \tilde{L}^2 \leq 1 + o(1)$ , which with  $\mathbb{E}_0 \tilde{L} \rightarrow 1$  shows that  $\text{Var}_0(\tilde{L}) \rightarrow 0$ . Let  $S_1, S_2 \stackrel{\text{iid}}{\sim} \pi$  and define  $K = |S_1 \cap S_2|$ . By Fubini's theorem, we have

$$\begin{aligned} \mathbb{E}_0 \tilde{L}^2 &= \mathbb{E}_{S_1, S_2} \mathbb{E}_0 \left( L_{S_1} L_{S_2} \mathbb{1}_{\Gamma_{S_1}} \mathbb{1}_{\Gamma_{S_2}} \right) \\ &= \pi^{\otimes 2} \left[ \mathbb{E}_0 \left( \exp \left( \theta(W_{S_1} + W_{S_2}) - 2\Lambda(\theta)n^{(2)} \right) \mathbb{1}_{\Gamma_{S_1} \cap \Gamma_{S_2}} \right) \right]. \end{aligned}$$

Define

$$W_{S \times T} = \frac{1}{2} \sum_{i \in S, j \in T} W_{i,j},$$

and note that  $W_S = W_{S \times S}$ . We use the decomposition

$$W_{S_1} + W_{S_2} = W_{S_1 \times (S_1 \setminus S_2)} + W_{S_2 \times (S_2 \setminus S_1)} + 2W_{S_1 \cap S_2}, \quad (49)$$

the fact that

$$\Gamma_{S_1} \cap \Gamma_{S_2} \subset \{W_{S_1 \cap S_2} \leq w_K\},$$

and the independence of the random variables on the RHS of (49), to get

$$\mathbb{E}_0 \left( \exp \left( \theta(W_{S_1} + W_{S_2}) - 2\Lambda(\theta)n^{(2)} \right) \mathbb{1}_{\Gamma_{S_1} \cap \Gamma_{S_2}} \right) \leq \text{I} \cdot \text{II} \cdot \text{III},$$

where

$$\text{I} := \mathbb{E}_0 \exp \left( \theta W_{S_1 \times (S_1 \setminus S_2)} - \frac{\Lambda(\theta)}{2} (n-K)(n+K-1) \right) = 1,$$

$$\begin{aligned} \text{II} &:= \mathbb{E}_0 \exp \left( \theta W_{S_2 \times (S_2 \setminus S_1)} - \frac{\Lambda(\theta)}{2} (n - K)(n + K - 1) \right) = 1, \\ \text{III} &:= \mathbb{E}_0 \left( \exp \left( 2\theta W_{S_1 \cap S_2} - 2\Lambda(\theta) K^{(2)} \right) \mathbb{1}_{\{W_{S_1 \cap S_2} \leq w_K\}} \right). \end{aligned}$$

The first two equalities are due to the fact that the likelihood integrates to one.

To bound III, we follow [Butucea and Ingster \(2011\)](#), with a twist. When  $K \leq k_{\min}$ , we will use the obvious bound:

$$\text{III} \leq \mathbb{E}_0 \exp \left( 2\theta W_{S_1 \cap S_2} - 2\Lambda(\theta) K^{(2)} \right) = \exp \left( \Delta K^{(2)} \right),$$

where

$$\Delta := \Lambda(2\theta) - 2\Lambda(\theta) = \log \left( 1 + \frac{(p_1 - p_0)^2}{p_0(1 - p_0)} \right). \quad (50)$$

When  $K > k_{\min}$ , we use a different bound. For any  $\xi \in (0, 2\theta)$ , we have

$$\begin{aligned} \text{III} &\leq \mathbb{E}_0 \left[ \exp \left( \xi W_{S_1 \cap S_2} + (2\theta - \xi) w_K - 2\Lambda(\theta) K^{(2)} \right) \mathbb{1}_{\{W_{S_1 \cap S_2} \leq w_K\}} \right] \\ &\leq \mathbb{E}_0 \exp \left[ \xi W_{S_1 \cap S_2} + (2\theta - \xi) w_K - 2\Lambda(\theta) K^{(2)} \right], \end{aligned}$$

so that

$$\text{III} \leq \exp \left( \Delta_K K^{(2)} \right),$$

where

$$\Delta_k := \min_{\xi \in [0, 2\theta]} \Lambda(\xi) + (2\theta - \xi) q_k - 2\Lambda(\theta). \quad (51)$$

By the variational definition of the entropy (39), the minimum of  $\Lambda(\xi) + (2\theta - \xi) q_k - 2\Lambda(\theta)$  over  $\xi$  in  $\mathbb{R}^+$  is achieved at  $\xi = \theta_{q_k}$ , and we know from Lemma 7 that  $\theta_{q_k} \leq 2\theta$ . Hence, we have

$$\begin{aligned} \Delta_k &= -H(q_k) + 2\theta q_k - 2\Lambda(\theta) \\ &= -2H_{p_1}(q_k) + H(q_k), \end{aligned} \quad (52)$$

Following our tracks, we have

$$\mathbb{E}_0 \tilde{L}^2 \leq \mathbb{E} \left[ \mathbb{1}_{\{K \leq k_{\min}\}} \exp \left( \Delta K^{(2)} \right) \right] + \mathbb{E} \left[ \mathbb{1}_{\{K > k_{\min}\}} \exp \left( \Delta_K K^{(2)} \right) \right],$$

where the expectation is with respect to  $\pi^{\otimes 2}$ .

Let  $b$  be an integer sequence such that  $b \rightarrow \infty$  so slowly that

$$\frac{(p_1 - p_0)}{\sqrt{p_0}} \frac{bn^2}{N} \rightarrow 0, \quad (53)$$

which is possible because of (7). Recall that  $\rho = n/(N - n)$  and define  $k_0 = \lceil bn\rho \rceil$ . We divide the expectation into two parts:  $K \leq k_0$  and  $k_0 + 1 \leq K \leq n$ . When  $k_0 = 1$ , we simply have

$$\mathbb{E} \left[ \mathbb{1}_{\{K \leq k_0\}} \exp \left( \Delta K^{(2)} \right) \right] = \mathbb{P}(K \leq 1) \leq 1.$$

When  $k_0 \geq 2$ , we use the expression (50) of  $\Delta$  to derive

$$\begin{aligned} \mathbb{E} \left[ \mathbb{1}_{\{K \leq k_0\}} \exp \left( \Delta K^{(2)} \right) \right] &\leq \exp \left[ \Delta k_0^2 \right] \\ &\leq \exp \left[ O(1) \frac{(p_1 - p_0)^2}{p_0(1 - p_0)} \frac{b^2 n^2}{N^2} \right] = 1 + o(1) \end{aligned}$$

because of (53).

When  $k_0 + 1 \leq K \leq \lfloor k_{\min} \rfloor$ , we use the bound (34) and the identity  $(1-x)\log(1-x) \geq -x$ , to get

$$\begin{aligned} \mathbb{E} \left[ \mathbb{1}_{\{k_0+1 \leq K \leq \lfloor k_{\min} \rfloor\}} \exp \left( \Delta K^{(2)} \right) \right] &\leq \sum_{k=k_0+1}^{\lfloor k_{\min} \rfloor} \exp \left[ \Delta \frac{k(k-1)}{2} - n H_\rho \left( \frac{k}{n} \right) \right] \\ &\leq \sum_{k=k_0+1}^{\lfloor k_{\min} \rfloor} \exp \left[ k \left( \Delta \frac{k-1}{2} - \log \left( \frac{k}{n\rho} \right) + 1 \right) \right] \end{aligned}$$

For  $a > 0$  fixed, the function  $f(x) = ax - \log x$  is decreasing on  $(0, 1/a)$  and increasing on  $(1/a, \infty)$ . Therefore, for  $k_0 + 1 \leq k \leq n$ ,

$$\Delta \frac{k-1}{2} - \log \left( \frac{k}{n\rho} \right) \leq -\omega ,$$

where

$$\omega := \min \left[ \log b - \Delta \frac{k_0-1}{2}, \log \left( \frac{k_{\min}}{n\rho} \right) - \Delta \frac{k_{\min}-1}{2} \right] .$$

From what we did previously, we know that  $\Delta(k_0-1) = o(1)$ , so that the first term in the maximum tends to  $\infty$ . Therefore, it suffices to look at the second term in the maximum. In fact,  $k_{\min}$  has been precisely defined in (45) to make this second term diverge. Indeed, by (45) and (50), we have

$$\Delta \frac{k_{\min}-1}{2} \leq \log \left( \frac{Nk^*}{n^2} \right) - \log \log \left[ \frac{n}{\log(N/n)} \right] .$$

By Lemma 6 and since  $\rho \asymp n/N = o(1)$ , we get  $\log(k_{\min}/(n\rho)) - \log \left( \frac{Nk^*}{n^2} \right) = o(1)$ . Consequently,

$$\log \left( \frac{k_{\min}}{n\rho} \right) - \Delta \frac{k_{\min}-1}{2} \geq \log \log \left[ \frac{n}{\log(N/n)} \right] + o(1) \rightarrow \infty ,$$

because of (5).

When  $K > k_{\min}$ , we have

$$\mathbb{E} \left[ \mathbb{1}_{\{K > k_{\min}\}} \exp \left( \Delta_K K^{(2)} \right) \right] \leq \sum_{k=\lfloor k_{\min} \rfloor + 1}^n \exp \left[ k \left( \Delta_k \frac{k-1}{2} - \log \left( \frac{k}{n\rho} \right) + 1 \right) \right] .$$

Now, using (52), we have

$$\Delta_k \frac{k-1}{2} - \log \left( \frac{k}{n\rho} \right) = \frac{k-1}{2} [-2H_{p_1}(q_k) + H(q_k)] - \log \left( \frac{N}{k} \right) + 2 \log \left( \frac{n}{k} \right) + o(1) ,$$

which goes to  $-\infty$  uniformly over all  $k$  between  $\lfloor k_{\min} \rfloor + 1$  and  $n$  by the definition (47) of  $q_k$  and by the control of  $H_{p_1}(q_k)$  from Lemma 8. Hence, the sum above tends to zero.

This concludes the proof that  $\mathbb{E}_0 \tilde{L}^2 \leq 1 + o(1)$ .

### 7.3.3 Proof of Lemma 6

We only need to prove that  $k_* \rightarrow \infty$  and that  $\log(n/k_*) = o[\log(N/n)]$  since

$$\log \left\{ \log \left( \frac{n}{\log(N/n)} \right) \wedge \log(N/n) \right\} = o(\log(N/n)) .$$

We divide the analysis into three cases depending on the behaviour of  $p_1/p_0$ .

**CASE 1:**  $p_1/p_0 \rightarrow 1$ . Then, Lemma 3 tells us  $\log \left( 1 + \frac{(p_1-p_0)^2}{p_0(1-p_0)} \right) \sim 2H(p_1)$ , so that

$$k^* \succ \frac{\log(N/n)}{H(p_1)} \wedge n \succ n ,$$

since  $H(p_1) < 2(1 - \eta_0) \log(N/n)/n$  by (40). Hence  $k^* \rightarrow \infty$  and  $\log(n/k^*) = O(1)$ .

**CASE 2:**  $p_1/p_0 \rightarrow r$  with  $r \in (1, \infty)$ . Since  $H(p_1)$  goes to 0, this enforces  $p_0 \rightarrow 0$ . Using Lemma 3 and (40), we derive that

$$p_0 [r \log(r) - r + 1] \prec \log(N/n)/n .$$

Hence,  $\log(N/n)/p_0 \succ n$ . Going back to the definition of  $k_*$ , we derive that

$$k^* \succ \left[ 1 + \frac{\log(N/n)}{p_0(r-1)^2} \right] \wedge n \succ n .$$

**CASE 3:**  $p_1/p_0 \rightarrow \infty$ . Again, we have  $p_0 \rightarrow 0$ . By Lemma 3 and (40),

$$p_1 \log \left( \frac{p_1}{p_0} \right) \prec \frac{\log(N/n)}{n} . \quad (54)$$

Hence,

$$\log \left( \frac{p_1}{p_0} \right) \prec \log [\log(N/n)/(np_0)] = o[\log(N/n)],$$

where the last part comes from (1). Hence,

$$k^* \succ \frac{\log(N/n)}{\log(p_1/p_0)} \rightarrow \infty .$$

Since (54) also implies that  $p_1 \prec \log(N/n)/n$ , we have

$$\frac{n}{k^*} \prec \frac{n \log(1 + p_1^2/p_0)}{\log(N/n)} \vee 1 \prec \frac{np_1^2}{\log(N/n)p_0} \vee 1 \prec \frac{\log(N/n)}{np_0} \vee 1 ,$$

so that  $\log(n/k^*) \leq \log [\log(N/n)/(np_0)] \vee 0 + O(1) = o[\log(N/n)]$  by (1).

### 7.3.4 Proof of Lemma 7

Define  $\tilde{q}$  by the equation

$$\frac{\tilde{q}}{1 - \tilde{q}} = \frac{p_1^2(1 - p_0)}{p_0(1 - p_1)^2} , \quad (55)$$

which implies  $\theta_{\tilde{q}} = 2\theta$ . Because  $H$  is strictly increasing and continuous on  $(p_0, \tilde{q})$ , to prove the existence of  $q_k$  it suffices to show that

$$\frac{k_{\min} - 1}{2} H(\tilde{q}) \geq \log(N/k_{\min}) + 2 .$$

As in the proof of the previous lemma, we consider different cases depending on the convergence of  $p_1/p_0$  and of  $p_1^2/p_0$ . In all cases, except the last one, we show that

$$k_* H(\tilde{q}) \geq 2(1 + \varepsilon) \log(N/n),$$

for some fixed  $\varepsilon > 0$ , which suffices by Lemma 6. If  $k_* < n$ , so that  $k_* \geq \frac{2}{\Delta} \log(N/n)$  (with  $\Delta$  defined in (50)). If  $k_* = n$  and  $k_{\min} < n$ , we have  $k_* \geq \frac{2}{\Delta} \log(N/n)(1 + o(1))$ . Hence, it is enough to prove that

$$H(\tilde{q}) \geq (1 + \varepsilon)\Delta, \quad \text{for some fixed } \varepsilon > 0.$$

The last case, Case 3(c) below — which corresponds to  $p_0 = o(\log(N/n)/n)$  and  $\log(n) = o(\log(N))$  — requires a more delicate treatment.

**CASE 1:**  $p_1/p_0 \rightarrow 1$ . By the definition of  $\tilde{q}$ , we have  $\tilde{q} - p_0 = (p_1 - p_0) \left[ 1 + \frac{p_1(1-p_1)}{p_0 - 2p_0p_1 + p_1^2} \right] \sim 2(p_1 - p_0)$  and Lemma 3 tells us that

$$H(\tilde{q}) \sim \frac{2(p_1 - p_0)^2}{p_0(1 - p_0)} \geq 2\Delta.$$

**CASE 2:**  $p_1/p_0 \rightarrow r$  with  $r \in (1, \infty)$ . Note that this forces  $p_1 \rightarrow 0$ . Here (55) implies that  $\tilde{q}/p_0 \sim (p_1/p_0)^2$ , so that  $H(\tilde{q}) \sim p_0(r^2 \log(r^2) - r^2 + 1)$  by Lemma 3. At the same time,  $\Delta \sim p_0(r - 1)^2$ , so that

$$\frac{H(\tilde{q})}{\Delta} \sim \frac{r^2 \log(r^2) - r^2 + 1}{(r - 1)^2} = 1 + \frac{2r(r \log(r) - r + 1)}{(r - 1)^2} > 1.$$

**CASE 3(a):**  $p_1/p_0 \rightarrow \infty$  and  $p_1^2/p_0 \rightarrow 0$ . We have  $\tilde{q}/p_0 \sim (p_1/p_0)^2 \rightarrow \infty$ , implying that  $H(\tilde{q}) \sim \tilde{q} \log(\tilde{q}/p_0) \sim 2(p_1^2/p_0) \log(p_1/p_0)$  by Lemma 3. Also,  $\Delta \sim \log(1 + p_1^2/p_0) \sim \frac{p_1^2}{p_0}$ . Hence,  $H(\tilde{q}) \gg \Delta$ .

**CASE 3(b):**  $p_1/p_0 \rightarrow \infty$  and  $p_1^2/p_0 \rightarrow r_2 \in (0, \infty)$ . Here  $\tilde{q} \rightarrow 1/(1 + r_2)$ , so that  $\tilde{q}/p_0 \rightarrow \infty$ , implying that  $H(\tilde{q}) \sim \tilde{q} \log(\tilde{q}/p_0) \asymp \log(1/p_0) \rightarrow \infty$ . Also,  $\Delta \rightarrow \log(1 + r_2)$ . Hence,  $H(\tilde{q}) \gg \Delta$ .

**CASE 3(c):**  $p_1^2/p_0 \rightarrow \infty$ . By Definition (44) of  $k_*$ , this implies  $k_* < n$ . By definition of  $\tilde{q}$ , we have  $\tilde{q} = 1 - o(1)$ , so that  $H(\tilde{q}) \sim \log(1/p_0)$ . On the other hand,  $\Delta \sim \log(p_1^2/p_0)$ . Therefore,

$$\frac{H(\tilde{q})}{\Delta} \sim \frac{\log(1/p_0)}{\log(p_1^2/p_0)} = \frac{1}{1 - \frac{\log(p_1^2)}{\log(p_0)}},$$

so that we are done if  $\log(p_1)/\log(p_0)$  is bounded away from 0. When  $\log(p_1)/\log(p_0) = o(1)$ , we need to work a little harder and perform a second order analysis. From the definition of  $\tilde{q}$ , we derive  $1 - \tilde{q} \leq \frac{p_0}{p_1^2}$ , so that

$$H(\tilde{q}) \geq H(1 - \frac{p_0}{p_1^2}) = (1 - \frac{p_0}{p_1^2}) \log(\frac{1 - \frac{p_0}{p_1^2}}{p_0}) + \frac{p_0}{p_1^2} \log(\frac{\frac{p_0}{p_1^2}}{1 - p_0}) = (1 - \frac{p_0}{p_1^2}) \log(\frac{1}{p_0}) + o(1).$$

Hence,

$$\begin{aligned} \frac{H(\tilde{q})}{\Delta} - 1 &\geq \frac{\log(1/p_1^2) - \frac{p_0}{p_1^2} \log(1/p_0) - o(1)}{\log\left(\frac{p_1^2}{p_0}\right) + o(1)} \\ &\geq \frac{2 \log(1/p_1)}{\log(1/p_0)} \left( \frac{1 - \frac{p_0 \log(p_0)}{p_1^2 \log(p_1^2)} + o(1)}{1 - \frac{2 \log(p_1)}{\log(p_0)} + o(1)} \right) \\ &= (2 + o(1)) \frac{\log(1/p_1)}{\log(1/p_0)}. \end{aligned}$$



since  $p_1^2/p_0 \rightarrow \infty$ . We use this lower bound to get

$$\begin{aligned} \frac{k_{\min} - 1}{2} H(\tilde{q}) &\geq [\log(N/k_*) - 2\log(n/k_*) - \log \log(n/\log(N/n))] \frac{H(\tilde{q})}{\Delta} \\ &\geq [\log(N/k_{\min}) + 2] \times \left[ 1 - \frac{2 + o(1) + 2\log(n/k_*) + \log \log(n/\log(N/n))}{\log(N/n)} \right] \\ &\quad \times \left[ 1 + (2 + o(1)) \frac{\log(1/p_1)}{\log(1/p_0)} \right], \end{aligned}$$

where we used Lemma 6 in the second inequality. In order to conclude, because of (5), it suffices to show that

$$\frac{\log(n/k_*) + \log \log(n/\log(N/n))}{\log(N/n)} \ll \frac{\log(1/p_1)}{\log(1/p_0)}. \quad (56)$$

The bound (54), coupled with  $p_1 \gg \sqrt{p_0}$ , implies that  $2\log \log(N/n) - \log(n) + \log(1/(np_0)) \rightarrow \infty$ . This, together with (1), forces  $\log(n) = o[\log(N/n)]$ . Hence,

$$\frac{\log(1/p_0)}{\log(N/n)} = \frac{\log(n) + \log(1/(np_0))}{\log(N/n)} = o(1).$$

It remains to show that

$$\frac{\log(n/k_*) + \log \log(n/\log(N/n))}{\log(1/p_1)} = O(1).$$

By definition of  $k_*$

$$\log(n/k_*) \leq \log(n/\log(N/n)) + \log(\Delta) \leq \log(n/\log(N/n)) + \log \log(p_1^2/p_0),$$

so that, because of (5) and (54), we have

$$\begin{aligned} \frac{\log(n/k_*) + \log \log(n/\log(N/n))}{\log(1/p_1)} &\prec \frac{\log(n/\log(N/n)) + \log \log(p_1^2/p_0) + \log \log(n/\log(N/n))}{\log(n/\log(N/n)) + \log \log(p_1/p_0)} \\ &= O(1). \end{aligned}$$

### 7.3.5 Proof of Lemma 8

We first note that, by the entropy bound (40) involving  $p_1$ , the definition of  $q_k$  Lemma 7, definition of  $\tilde{q}$  in (55), and the fact that  $H(q)$  is strictly increasing over  $q > p_0$ , we have

$$p_1 \leq q_k \leq \tilde{q}, \quad \forall k \leq n. \quad (57)$$

**CASE 1:**  $p_1/p_0 \rightarrow 1$ . In the proof of Lemma 7 (Case 1), we have shown that  $\tilde{q}$  defined in (55) satisfies  $\tilde{q} \sim p_0$ . By (57), we then get  $q_k \sim p_0 \sim p_1$ . Then using Lemma 3 and the bound on the entropy (40), we get

$$\frac{(q_k - p_0)^2}{(p_1 - p_0)^2} \sim \frac{H(q_k)}{H(p_1)} \geq \frac{n}{(1 - \eta_0)k} \geq \frac{1}{1 - \eta_0}. \quad (58)$$

Hence, we may lower bound  $H_{p_1}(q_k)$  as follows:

$$H_{p_1}(q_k) \sim \frac{(q_k - p_1)^2}{2p_1(1 - p_1)} \sim \frac{(q_k - p_0)^2}{2p_0(1 - p_0)} \left( 1 - \frac{p_1 - p_0}{q_k - p_0} \right)^2 \succ H(q_k)[1 - \sqrt{1 - \eta_0}]^2,$$

which allows us to conclude that

$$\frac{(k-1)}{2}H_{p_1}(q_k) \succ \frac{(k-1)}{2}H(q_k) \succ \log(N/n) \gg \log\left(\frac{n}{k}\right) \vee 1 ,$$

where the last inequality follows from Lemma 6 and the fact that  $k \geq k_{\min}$ .

**CASE 2:**  $p_1/p_0 \rightarrow r \in (1, \infty)$ . As in the proof of Lemma 7 (Case 2), we have  $p_1 \rightarrow 0$ . In the proof of Lemma 7 (Case 1), we have shown that  $\tilde{q}/p_0 \rightarrow r^2$  and that  $\tilde{q} \rightarrow 0$ . By (57), we can use the second asymptotic expression of the entropies in Lemma 3. The inequalities in (58) still hold, giving

$$\frac{1}{1-\eta_0} \leq \frac{H(q_k)}{H(p_1)} \sim \frac{\frac{q_k}{p_0} \log\left(\frac{q_k}{p_0}\right) - \frac{q_k}{p_0} + 1}{r \log(r) - r + 1} = \frac{f(q_k/p_0)}{f(r)} , \quad (59)$$

where  $f(x) := x \log(x) - x + 1$ . Since  $f$  is convex and satisfies  $f'(x) = \log(x)$ , we have  $f(x) - f(r) \leq (x - r) \log(x)$  for  $x \geq r \geq 1$ . Taking  $x = q_k/p_0$  and using (59), we derive that

$$\log\left(\frac{q_k}{p_0}\right) \left(\frac{q_k}{p_0} - r\right) \geq f(r) \left(\frac{f(q_k/p_0)}{f(r)} - 1\right) \geq \frac{f(r)\eta_0}{1-\eta_0}(1 + o(1)) \geq f(r)\eta_0 ,$$

eventually. As a consequence,  $q_k/p_0$  is also lower bounded away from  $r$ . Thus,  $\log(q_k/p_1)/\log(q_k/p_0)$  is bounded away from 0 by a constant that only depends on  $r$  and  $\eta_0$ . We then derive,

$$\log\left(\frac{q_k}{p_1}\right) \left(\frac{q_k}{p_1} - 1\right) \succ \log\left(\frac{q_k}{p_0}\right) \left(\frac{q_k}{p_0} - r\right) . \quad (60)$$

Now, for the entropy  $H_{p_1}(q_k)$ , by Lemma 3 we have

$$\begin{aligned} H_{p_1}(q_k) &\succ \frac{(q_k - p_1)^2}{p_1} \wedge q_k \log\left(\frac{q_k}{p_1}\right) \\ &= p_1 \left[ \left(\frac{q_k}{p_1} - 1\right)^2 \wedge \frac{q_k}{p_1} \log\left(\frac{q_k}{p_1}\right) \right] \geq p_1 \left(\frac{q_k}{p_1} - 1\right) \log\left(\frac{q_k}{p_1}\right) \end{aligned}$$

as  $\log(1+x) \leq x$ . Since  $H(p_1) \sim p_0 f(r)$ , we get by (59) and (60)

$$\begin{aligned} H_{p_1}(q_k) &\succ \frac{rH(p_1)}{f(r)} \left(\frac{q_k}{p_1} - 1\right) \log\left(\frac{q_k}{p_1}\right) \\ &\succ \frac{H(q_k)}{f(q_k/p_0)} \left(\frac{q_k}{p_0} - r\right) \log\left(\frac{q_k}{p_0}\right) \\ &\succ H(q_k) \\ &\succ \frac{1}{k} \log(N/n) , \end{aligned}$$

where the third line follows from the fact that the  $q_k/p_0$  is lower bounded away from  $r$  and that  $f(x) \sim x \log(x)$  when  $x \rightarrow \infty$ . Thus,

$$\frac{k-1}{2}H_{p_1}(q_k) \succ \log(N/n) \gg \log(n/k) \vee 1 ,$$

as before.

**CASE 3:**  $p_1/p_0 \rightarrow \infty$ . As in the proof of Lemma 7 (Case 2), we have  $p_1 \rightarrow 0$ . We start as in the two previous cases, again using Lemma 3 to get the asymptotic expressions of the entropies. By (57),  $q_k/p_0 \geq p_1/p_0 \rightarrow \infty$ , so that

$$\frac{n}{(1-\eta_0)(k-1)} \leq \frac{H(q_k)}{H(p_1)} \sim \frac{q_k \log(q_k/p_0)}{p_1 \log(p_1/p_0)} \sim \frac{q_k}{p_1} \left[ 1 + \frac{\log(q_k/p_1)}{\log(p_1/p_0)} \right] \quad (61)$$

It follows that  $\frac{q_k}{p_1} (1 + \frac{\log(q_k/p_1)}{\log(p_1/p_0)}) \geq (1-\eta_0)^{-1}$ . Since  $\log(p_1/p_0) \rightarrow \infty$ , we derive that  $q_k/p_1 \geq (1-\eta_0/2)^{-1}$  for  $n$  large enough. Since  $p_1 \leq q_k \leq \tilde{q}$ , we have  $q_k/p_1 \leq \tilde{q}/p_1 \leq p_1/p_0$ . It follows that  $\log(q_k/p_1)/\log(p_1/p_0) \leq 1$ , and therefore  $q_k/p_1 \geq (1+o(1))\frac{n}{2k}$  by (61). We conclude that

$$\frac{q_k}{p_1} \geq \left[ \frac{n}{2k} \vee \frac{1}{1-\eta_0/2} \right] (1+o(1)) . \quad (62)$$

Turning to the entropy  $H_{p_1}(q_k)$ , we have  $H_{p_1}(q_k) \geq q_k \log(q_k/p_1) - q_k + (1-q_k)p_1$ . Using Lemma 7 and Lemma 3, we get

$$\frac{k-1}{2} H_{p_1}(q_k) \geq \frac{H_{p_1}(q_k)}{H_{p_0}(q_k)} \log\left(\frac{N}{k}\right) \geq \frac{\log\left(\frac{q_k}{p_1}\right) - 1 + \frac{p_1}{q_k} - p_1}{\log\left(\frac{q_k}{p_0}\right)} \log\left(\frac{N}{n}\right) (1+o(1)) .$$

We explain above that  $q_k/p_1 \leq \tilde{q}/p_1 \leq p_1/p_0$ , so that  $q_k/p_0 \leq (p_1/p_0)^2$ , implying  $\log(q_k/p_0) \leq 2\log(p_1/p_0)$ . Applying (62), we get

$$\frac{k-1}{2} H_{p_1}(q_k) \succcurlyeq \frac{\log(N/n)}{\log(p_1/p_0)} [\log(n/k) \vee 1] ,$$

We saw in the proof of Lemma 6 (Case 3) that  $\log(p_1/p_0) = o[\log(N/n)]$ , so we conclude that

$$\frac{k-1}{2} H_{p_1}(q_k) \gg \log(n/k) \vee 1 .$$

## 7.4 Proof of Theorem 3

We start with a couple of lemmas.

**Lemma 9.** *Under conditions (17), (18) and (19), we have*

$$\limsup \frac{nH_{p_0}(p_1)}{2\log(N/n)} < 1 , \quad \frac{(p_1-p_0)^2}{p_0} \frac{n^3}{N^{3/2}} \rightarrow 0 . \quad (63)$$

As in the proof of Theorem 2, for  $n$  large enough, we may assume that there exists  $\eta_0 > 0$  such that

$$\frac{nH_{p_0}(p_1)}{2\log(N/n)} = 1 - \eta_0 . \quad (64)$$

**Lemma 10.** *Under conditions (17) and (64), we have*

$$\frac{n^2}{N} \frac{(p_1-p_0)^2}{p_0} = o(1) .$$

We consider the likelihood ratio under the uniform prior:

$$L' = \binom{N}{n}^{-1} \sum_{|S|=n} L'_S = \pi[L'_S], \quad (65)$$

and

$$L'_S := \exp \left[ \theta_{p_1} W_S - \Lambda(\theta_{p_1}) n^{(2)} + \theta_{p'_0} (W - W_S) - (N^{(2)} - n^{(2)}) \Lambda(\theta_{p'_0}) \right]. \quad (66)$$

As in the proof of Theorem 2, we use a thresholded version of  $L'$  to prove that  $\mathbb{E}_0[|L' - 1|] = o(1)$ :

$$\tilde{L} := \binom{N}{n}^{-1} \sum_{|S|=n} L'_S \mathbf{1}_{\Gamma_S},$$

where  $\Gamma_S$  is defined in (46). As in the proof of Theorem 2, we prove that any subsequence of  $\mathbb{E}_0[\tilde{L} - 1]$  has 0 as an accumulation point. This allows us to assume that  $p_1/p_0$  converges to  $r \in [1, \infty]$  and that  $p_1^2/p_0$  converges to  $r_2 \in [0, \infty]$ . To control  $\mathbb{E}_0[\tilde{L} - 1]$ , it suffices to prove that  $\mathbb{E}_0 \tilde{L} = 1 + o(1)$  and that  $\mathbb{E}_0[\tilde{L}^2] \leq 1 + o(1)$ .

### First moment

$$\mathbb{E}_0 \tilde{L} = \pi[\mathbb{E}_0[L'_S \mathbf{1}_{\Gamma_S}]] = \pi[\mathbb{P}'_S(\Gamma_S)] = \mathbb{P}'_S(\Gamma_S).$$

As the proof of Theorem 2, we can show that  $\mathbb{P}'_S(\Gamma_S) = 1 + o(1)$  relying only on (19).

**Second Moment.** It remains to prove that  $\mathbb{E}_0[\tilde{L}^2] \leq 1 + o(1)$ . Let  $S_1, S_2 \stackrel{\text{iid}}{\sim} \pi$  and define  $K = |S_1 \cap S_2|$ . Observe that  $(W_{S_1 \cap S_2}, W_{S_1} + W_{S_2} - 2W_{S_1 \cap S_2}, W - W_{S_1} - W_{S_2} + W_{S_1 \cap S_2})$  are independent. Arguing as in the proof of Theorem 2, we decompose the square of the modified likelihood as follows.

$$\begin{aligned} \mathbb{E}_0 \tilde{L}^2 &= \pi^{\otimes 2}[\mathbb{E}_0(L'_{S_1} L'_{S_2} \mathbf{1}_{\Gamma_{S_1}} \mathbf{1}_{\Gamma_{S_2}})] \\ &\leq \pi^{\otimes 2}[\text{I} \cdot \text{II} \cdot \text{III}] \end{aligned}$$

where

$$\begin{aligned} \text{I} &:= \mathbb{E}_0 \exp \left[ 2\theta_{p'_0} (W - W_{S_1} - W_{S_2} + W_{S_1 \cap S_2}) - 2\Lambda(\theta_{p'_0}) (N^{(2)} - 2n^{(2)} + K^{(2)}) \right], \\ \text{II} &:= \mathbb{E}_0 \exp \left[ \left( \theta_{p_1} + \theta_{p'_0} \right) (W_{S_1} + W_{S_2} - 2W_{S_1 \cap S_2}) - 2 \left( \Lambda(\theta_{p_1}) + \Lambda(\theta_{p'_0}) \right) (n^{(2)} - K^{(2)}) \right], \\ \text{III} &:= \mathbb{E}_0 \left[ \exp \left( 2\theta_{p_1} W_{S_1 \cap S_2} - 2\Lambda(\theta_{p_1}) K^{(2)} \right) \mathbf{1}_{\{W_{S_1 \cap S_2} \leq w_K\}} \right]. \end{aligned}$$

All these expectations only depend on  $S_1$  and  $S_2$  through  $K$ .

The term III already appeared in the proof of Theorem 2, where we saw that  $\text{III} \leq \exp(\Delta K^{(2)})$  for  $K \leq k_{\min}$ , and that  $\text{III} \leq \exp(\Delta_K K^{(2)})$  for  $K > k_{\min}$  where  $k_{\min}$  is defined in (45), while  $\Delta$  and  $\Delta_K$  are defined in (50) and (51), respectively.

Since the expectations inside I and II are not thresholded, we easily compute these terms:

$$\text{I} = \exp \left[ \left( N^{(2)} - 2n^{(2)} + K^{(2)} \right) (\Lambda(2\theta_{p'_0}) - 2\Lambda(\theta_{p'_0})) \right],$$

with

$$\Lambda(2\theta_{p'_0}) - 2\Lambda(\theta_{p'_0}) = \log \left( 1 + \frac{(p'_0 - p_0)^2}{p_0(1 - p_0)} \right) \leq \frac{(p_1 - p'_0)^2}{p_0(1 - p_0)} \left( \frac{n^{(2)}}{N^{(2)}} \right)^2;$$

and

$$\Pi = \exp \left[ 2 \left( n^{(2)} - K^{(2)} \right) \left( \Lambda(\theta_{p_1} + \theta_{p'_0}) - \Lambda(\theta_{p_1}) - \Lambda(\theta_{p'_0}) \right) \right] ,$$

with

$$\Lambda(\theta_{p_1} + \theta_{p'_0}) - \Lambda(\theta_{p_1}) - \Lambda(\theta_{p'_0}) = \log \left( 1 - \frac{(p_0 - p'_0)(p_1 - p_0)}{p_0(1 - p_0)} \right) \leq -\frac{(p_1 - p'_0)(p_1 - p_0)}{p_0(1 - p_0)} \frac{n^{(2)}}{N^{(2)}} .$$

Since  $(p_1 - p'_0) = (p_1 - p_0)(1 - n^{(2)}/N^{(2)})^{-1}$ , we derive

$$\text{I} \cdot \Pi \leq \exp \left[ \frac{(p_1 - p_0)^2}{p_0(1 - p_0)} \left( -\frac{(n^{(2)})^2}{N^{(2)}} + \frac{n^{(2)}}{N^{(2)}} \left( K^{(2)} - \frac{(n^{(2)})^2}{N^{(2)}} \right) \frac{2 - \frac{n^{(2)}}{N^{(2)}}}{\left( 1 - \frac{n^{(2)}}{N^{(2)}} \right)^2} \right) \right] =: V_K . \quad (67)$$

By Lemma 10,  $\Delta n^2/N \rightarrow 0$  and by (63),  $\Delta n^3/N^{3/2} \rightarrow 0$ . Hence, there exists  $b \rightarrow \infty$  such that  $\Delta \frac{n^3}{N^{3/2}} b^2 \rightarrow 0$  and  $\Delta b \frac{n^2}{N} \rightarrow 0$ . Define  $k'_0 = \lfloor \frac{n^2}{N} + \frac{n}{N^{1/2}} b \rfloor$  and  $k_0 = \lfloor b \frac{n^2}{N} \rfloor$ . We can take  $b$  small enough to constrain  $k_0 \leq n/2$ .

To prove that  $\mathbb{E}_0 \tilde{L}^2 \leq 1 + o(1)$ , we only need to show the four following results

$$\mathbb{E} \left[ \{K \leq k'_0\} \exp \left\{ \Delta K^{(2)} \right\} V_K \right] \leq 1 + o(1) , \quad (68)$$

$$\mathbb{E} \left[ \{k'_0 < K \leq k_0\} \exp \left\{ \Delta K^{(2)} \right\} V_K \right] = o(1) , \quad (69)$$

$$\mathbb{E} \left[ \{k_0 < K \leq k_{\min}\} \exp \left\{ \Delta K^{(2)} \right\} V_K \right] = o(1) . \quad (70)$$

$$\mathbb{E} \left[ \{k_{\min} < K \leq n\} \exp \left\{ \Delta_K K^{(2)} \right\} V_K \right] = o(1) . \quad (71)$$

By Lemma 10 and the definition (67) of  $V_k$ , we have  $\log(V_k) = o(k^2/N) = o(k)$  when  $k \leq n$ . As a consequence, the expectations in (70) and (71) are almost the same as the expectations  $\mathbb{E} [\{k_0 < K \leq k_{\min}\} \exp \{ \Delta K^{(2)} \}]$  and  $\mathbb{E} [\{k_{\min} < K \leq n\} \exp \{ \Delta_K K^{(2)} \}]$  that we bounded in the proof of Theorem 2. This is made rigorous to establish the following result.

**Lemma 11.** *Under the entropy condition (64), the bounds (70) and (71) hold.*

In fact the main difference between the proof of Theorem 2 and the current proof lies in the control of the two expectations in (68) and (69). Here, we need to carefully upper bound  $V_K$  in order to balance  $\Delta K^{(2)}$ . Using the identity  $\log(1 + x) \leq x$ , the property  $\log(V_k) \leq 0$  for  $k \leq n/2$  — easily verified from the definition (67) — and  $k_0 \leq n/2$ , we get

$$V_k \leq \exp \left[ \Delta \left( -\frac{(n^{(2)})^2}{N^{(2)}} + \frac{n^{(2)}}{N^{(2)}} \left( k^{(2)} - \frac{(n^{(2)})^2}{N^{(2)}} \right) \frac{2 - \frac{n^{(2)}}{N^{(2)}}}{\left( 1 - \frac{n^{(2)}}{N^{(2)}} \right)^2} \right) \right] ,$$

for  $k \leq k_0$ . In the sequel, we note

$$\Delta' := \frac{\Delta}{2} \left[ 1 + \frac{n^{(2)}}{N^{(2)}} \frac{2 - \frac{n^{(2)}}{N^{(2)}}}{\left( 1 - \frac{n^{(2)}}{N^{(2)}} \right)^2} \right] ,$$

so that  $2\Delta' \sim \Delta$ . Thus, we get for any  $k \leq k_0$ ,

$$\begin{aligned} \exp \left\{ \Delta k^{(2)} \right\} V_k &\leq \exp \left\{ 2\Delta' \left( k^{(2)} - \frac{(n^{(2)})^2}{N^{(2)}} \right) \right\} \\ &\leq \exp \left\{ \Delta' \left( k^2 - \frac{n^4}{N^2} \right) + 2\Delta' \frac{n^3}{N^2} \right\} \\ &\leq (1 + o(1)) \exp \left\{ \Delta' \left( k^2 - \frac{n^4}{N^2} \right) \right\}, \end{aligned} \quad (72)$$

since  $\Delta' n^3/N^2 \prec \Delta n^3/N^2 \leq \frac{(p_1-p_0)^2}{p_0(1-p_0)} \frac{n^3}{N^2} = o(n/N) = o(1)$  by Lemma 10.

Using this upper bound (72), we consider the expectation in (68)

$$\begin{aligned} \mathbb{E} \left[ \{K \leq k'_0\} \exp \left\{ \Delta K^{(2)} \right\} V_K \right] &\prec \exp \left[ \Delta' \left( k_0'^2 - \frac{n^4}{N^2} \right) \right] \\ &\prec \exp \left[ \Delta' \left( k'_0 - \frac{n^2}{N} \right) \left( k'_0 + \frac{n^2}{N} \right) \right] \\ &\leq \exp \left[ \Delta' \frac{bn}{N^{1/2}} \left( \frac{2n^2}{N} + \frac{bn}{N^{1/2}} \right) \right] = 1 + o(1) \end{aligned}$$

since  $\Delta' \frac{b^2 n^2}{N} \prec \Delta \frac{b^2 n^2}{N} = o(1)$  and  $\Delta' \frac{bn^3}{N^{3/2}} \ll \Delta \frac{b^2 n^3}{N^{3/2}} = o(1)$  by definition of  $b$ . We have proved (68).

To prove (69), we apply the Cauchy-Schwarz inequality and we upper bound  $K$  by  $k_0 \leq bn^2/N$ ,

$$\begin{aligned} \mathbb{E} \left[ \{k'_0 < K \leq k_0\} \exp \left\{ \Delta' \left( K^2 - \frac{n^4}{N^2} \right) \right\} \right] &\leq \mathbb{E} \left[ \{k'_0 < K \leq k_0\} \exp \left\{ \Delta' (b+1) \frac{n^2}{N} \left( K - \frac{n^2}{N} \right) \right\} \right] \\ &\leq \mathbb{P}^{1/2}(K > k'_0) \mathbb{E}^{1/2} \left[ \exp \left\{ 2\Delta' (b+1) \frac{n^2}{N} \left( K - \frac{n^2}{N} \right) \right\} \right] \end{aligned}$$

Recall that  $K \sim \text{Hyp}(N, n, n)$ , so that  $\mathbb{E} K = \frac{n^2}{N}$  and  $\text{Var}(K) \leq \frac{n^2}{N}$ . Hence, by Chebyshev's inequality,  $\mathbb{P}(K > k'_0) \leq 1/b^2 \rightarrow 0$ .

We know from (Aldous, 1985, p.173) that  $K$  has the same distribution as the random variable  $\mathbb{E}(W|\mathcal{B}_p)$  where  $W$  is binomial random variable of parameters  $n$ ,  $n/N$  and  $\mathcal{B}_N$  some suitable  $\sigma$ -algebra. By a convexity argument, we apply this to get

$$\begin{aligned} \mathbb{E} \exp \left\{ 2\Delta' (b+1) \frac{n^2}{N} \left( K - \frac{n^2}{N} \right) \right\} &\leq \left[ 1 + \frac{n}{N} \left( e^{2\Delta' (b+1) \frac{n^2}{N}} - 1 \right) \right]^n e^{-2\Delta' (b+1) \frac{n^4}{N^2}} \\ &\leq \exp \left[ 4 \frac{n^6}{N^3} (b+1)^2 \Delta'^2 \right] \leq 1 + o(1), \end{aligned}$$

since  $\Delta' b(\frac{n^2}{N} \vee \frac{n^3}{N^{3/2}}) = o(1)$  by definition of  $b$ . All in all, we have proved (69).

#### 7.4.1 Proof of Lemma 9

The second convergence is a straightforward consequence of the definition of  $p_0$ , (18) and (19), so that we focus on the first result. Let us compute the difference between the two entropies  $H_{p'_0}(p_1)$

and  $H_{p_0}(p_1)$ .

$$\begin{aligned}
H_{p'_0}(p_1) - H_{p_0}(p_1) &= p_1 \log \left( \frac{p_0}{p'_0} \right) + (1 - p_1) \log \left( \frac{1 - p_0}{1 - p'_0} \right) \\
&\leq \frac{n^{(2)}}{N^{(2)}} \left[ p_1 \frac{p_1 - p'_0}{p'_0} - (1 - p_1) \frac{p_1 - p'_0}{1 - p'_0} \right] \\
&\leq \frac{n^{(2)}}{N^{(2)}} \frac{(p_1 - p'_0)^2}{p'_0(1 - p'_0)}
\end{aligned}$$

Arguing as in the proof of Lemma 10, we note that, under conditions (17) and (64),

$$\frac{n^2 (p_1 - p'_0)^2}{N p'_0(1 - p'_0)} = o(1) ,$$

so that  $H_{p'_0}(p_1) - H_{p_0}(p_1) = o(1/N) = o(\log(N/n)/n)$ , since  $n \leq N$ .

#### 7.4.2 Proof of Lemma 10

**CASE 1:**  $p_1/p_0 \rightarrow 1$ . By condition (64),

$$\frac{n^2 (p_1 - p_0)^2}{N p_0(1 - p_0)} \sim 2H(p_1) \frac{n^2}{N} \prec \log \left( \frac{N}{n} \right) \frac{n}{N} = o(1) .$$

**CASE 2:**  $p_1/p_0 \rightarrow c \in (1, \infty)$ . Similarly,

$$\begin{aligned}
\frac{n (p_1 - p_0)^2}{N p_0(1 - p_0)} &\prec p_0(c - 1)^2 \frac{n^2}{N} \\
&\prec H(p_1) \frac{n^2}{N} \prec \log \left( \frac{N}{n} \right) \frac{n}{N} = o(1) .
\end{aligned}$$

**CASE 3:**  $p_1/p_0 \rightarrow \infty$ . We have

$$\frac{n^2 (p_1 - p_0)^2}{N p_0(1 - p_0)} \sim \frac{p_1^2 n^2}{p_0 N} .$$

By condition (64) and  $p_1 \log(p_1/p_0) \sim H(p_1) \prec \frac{1}{n} \log(N/n)$ . Dividing this inequality by  $p_0$  and then taking the logarithm leads to  $\log(p_1/p_0) \prec \log \log(N/n) + \log(1/n p_0) = o(\log(N/n))$  by (17). It follows that  $p_1/p_0 = o(\sqrt{N/n})$  and  $p_1 = o(\log(N/n)/n)$ . All in all, we conclude that

$$\frac{p_1^2 n^2}{p_0 N} = o \left[ \log(N/n) \sqrt{\frac{n}{N}} \right] = o(1) .$$

#### 7.4.3 Proof of Lemma 11

Let us first consider (71). Using the upper bound  $\log(V_k) = o(k)$ , we only have to prove that

$$\mathbb{E} \left[ \{K > k_{\min}\} \exp \left( \Delta_K K^{(2)} + o(K) \right) \right] = o(1) .$$

We have shown in the proof of Theorem 2 (only using the entropy condition) that

$$\mathbb{E} \left[ \{K \geq k_{\min}\} \exp \left( \Delta_K K^{(2)} \right) \right] \leq \sum_{k=\lfloor k_{\min} \rfloor + 1}^n \exp \left[ k \left( \Delta_k \frac{k-1}{2} - \log \left( \frac{k}{n\rho} \right) + 1 \right) \right]$$



tends to zero since all the terms  $\Delta_k \frac{k-1}{2} - \log \left( \frac{k}{n\rho} \right) + 1$  simultaneously go to  $-\infty$  for  $k = \lfloor k_{\min} \rfloor + 1, \dots, n$ . Consequently,

$$\mathbb{E} \left[ \{K > k_{\min}\} \exp \left( \Delta_K K^{(2)} + o(K) \right) \right] \leq \sum_{k=\lfloor k_{\min} \rfloor + 1}^n \exp \left[ k \left( \Delta_k \frac{k-1}{2} - \log \left( \frac{k}{n\rho} \right) + 1 + o(1) \right) \right]$$

also tends to zero.

Let us turn to (70) following again the same arguments as in the proof of Theorem 2.

$$\begin{aligned} & \mathbb{E} \left[ \{k_0 < K \leq \lfloor k_{\min} \rfloor\} \exp \left\{ \Delta K^{(2)} + o(K) \right\} \right] \\ & \leq \sum_{k=k_0+1}^{\lfloor k_{\min} \rfloor} \exp \left[ \Delta k^{(2)} + o(k) - nH_\rho \left( \frac{k}{n} \right) \right] \\ & \leq \sum_{k=k_0+1}^{\lfloor k_{\min} \rfloor} \exp \left[ k \left\{ \frac{\Delta}{2}(k-1) + o(1) - \log \left( \frac{k}{n\rho} \right) + 1 \right\} \right]. \end{aligned}$$

Hence, as in the previous proof, we only need to prove that

$$\omega := \min \left[ \log b - \Delta k_0/2, \log \left( \frac{k_{\min}}{n\rho} \right) - \Delta k_{\min}/2 \right]$$

goes to  $\infty$ . By definition of  $k_0$ , we have  $\Delta k_0 = o(1)$ , while we showed in the previous proof that  $\log \left( \frac{k_{\min}}{n\rho} \right) - \Delta k_{\min} \rightarrow \infty$ . With this, we conclude.

## 7.5 Proof of Proposition 2

We start with a useful result for proving that a test is asymptotically powerful based on the first two moments of the corresponding test statistic.

**Lemma 12.** *Suppose that for testing  $H_0$  versus  $H_1$ , a statistic  $T$  satisfies*

$$R_T := \frac{\mathbb{E}_1(T) - \mathbb{E}_0(T)}{\max(\sqrt{\text{Var}_1(T)}, \sqrt{\text{Var}_0(T)})} \rightarrow \infty. \quad (73)$$

*Then there is a test based on  $T$  that is asymptotically powerful.*

*Proof.* Consider the test that rejects when  $T \geq \mathbb{E}_0(T) + \sqrt{R_T \text{Var}_0(T)}$ . By Chebyshev's inequality, the probability of type I error tends to zero:

$$\mathbb{P}_0(T \geq \mathbb{E}_0(T) + \sqrt{R_T \text{Var}_0(T)}) \leq \frac{1}{R_T} \rightarrow 0.$$

For the probability of type II error, we have

$$\mathbb{P}_1(T \geq \mathbb{E}_0(T) + \sqrt{R_T \text{Var}_0(T)}) = \mathbb{P}_1 \left( \frac{T - \mathbb{E}_1(T)}{\sqrt{\text{Var}_1(T)}} \geq -\gamma \right) \geq 1 - \frac{1}{\gamma^2},$$

where

$$\gamma := \frac{R_T \max(\sqrt{\text{Var}_1(T)}, \sqrt{\text{Var}_0(T)}) - \sqrt{R_T \text{Var}_0(T)}}{\sqrt{\text{Var}_1(T)}} \rightarrow \infty.$$

□

We now apply Lemma 12 to the total degree test. From (12), under the null,

$$\mathbb{E}_0(W) = \frac{N(N-1)}{2}p_0, \quad \text{Var}_0(W) = \frac{N(N-1)}{2}p_0(1-p_0),$$

while under the alternative,

$$\mathbb{E}_1(W) = \frac{N(N-1)}{2}p_0 + \frac{n(n-1)}{2}(p_1-p_0),$$

and

$$\text{Var}_1(W) = \frac{N(N-1)}{2}p_0(1-p_0) + \frac{n(n-1)}{2}[p_1(1-p_1) - p_0(1-p_0)].$$

In any case,

$$\max(\text{Var}_1(W), \text{Var}_0(W)) \leq \frac{1}{2}N^2p_0 + \frac{1}{2}n^2(p_1-p_0).$$

Recalling the definition of  $R_W$  in (73), under (13) we have

$$R_W \geq \frac{n(n-1)(p_1-p_0)}{\sqrt{N^2p_0 + n^2(p_1-p_0)}} \asymp \frac{n^2}{N} \frac{p_1-p_0}{\sqrt{p_0}} \rightarrow \infty.$$

Therefore, the total degree test is powerful when (13) holds.

## 7.6 Proof of Proposition 3

We use the union bound, Chernoff's bound (28) and (30) to get

$$\begin{aligned} \mathbb{P}_0(W_{[n]}^* \geq an^{(2)}) &\leq \binom{N}{n} \exp(-n^{(2)}H(a)) \\ &\leq \exp\left(n \log(Ne/n) - n^{(2)}H(a)\right), \end{aligned}$$

which goes to zero when

$$\log(N/n) - \frac{(n-1)}{2}H(a) \rightarrow -\infty. \quad (74)$$

Choose  $a = \eta p_0 + (1-\eta)p_1$  with  $\eta \in (0,1)$  fixed, sufficiently small that

$$\liminf \frac{nH(a)}{2 \log(N/n)} > 1.$$

This is possible because of how  $H$  varies, which is described in Lemma 3. We then consider the test that rejects when  $W_{[n]}^* \geq an^{(2)}$ . We just chose  $a$  so that its level tends to zero. Under the alternative, let  $S$  denote the community. By definition,  $W_{[n]}^* \geq W_S$ , and since  $W_S \sim \text{Bin}(n^{(2)}, p_1)$  and  $p_1 n^{(2)} \rightarrow \infty$ ,  $W_S = p_1 n^{(2)} + O_P(\sqrt{p_1 n^{(2)}})$ . Therefore, the test is powerful when  $p_1 - a \gg \sqrt{p_1 n^{(2)}}$ . Since  $p_1 - a = \eta(p_1 - p_0)$  and  $\eta > 0$  is constant, this is the same as  $(p_1 - p_0)n^2 \gg \sqrt{p_1 n^2}$ . Now, if  $p_1/p_0$  is bounded away from 1, this is true because  $p_1 - p_0 \asymp p_1$  and  $p_1 n^2 \rightarrow \infty$ ; while if  $p_1/p_0 \rightarrow 1$ , we use Lemma 3 and (15) to get that  $(p_1 - p_0)^2 n/p_0 \geq \text{cst} \log(N/n)$ , implying that  $(p_1 - p_0)n^2/\sqrt{p_1 n^2} \sim (p_1 - p_0)n/p_0 \rightarrow \infty$ .

## 7.7 Proof of Proposition 4

The arguments are based on cumbersome, but pedestrian moment calculations.

**Under the null.** We first show that  $V^*$  remains bounded under the null. Rewrite  $V$  as

$$\begin{aligned} V &= \frac{1}{N-2} \sum_{i=1}^N (W_i - (N-1)p_0)^2 + (\hat{p}_0 - p_0)^2 (N-1) \left[ -\frac{N(N-1)}{N-2} + \frac{N^{(2)}}{N^{(2)}-1} \right] \\ &\quad + (\hat{p}_0 - p_0)(N-1) \frac{N^{(2)}}{N^{(2)}-1} (-1 + 2p_0) - (N-1) \frac{N^{(2)}}{N^{(2)}-1} p_0(1-p_0). \end{aligned} \quad (75)$$

Since  $\mathbb{E}_0(\hat{p}_0 - p_0)^2 = (N^{(2)})^{-1} p_0(1-p_0)$  and  $\mathbb{E}_0(W_i - (N-1)p_0)^2 = (N-1)p_0(1-p_0)$ , it follows that  $\mathbb{E}_0 V = 0$ . For the variance, we have

$$\begin{aligned} \text{Var}_0 [(\hat{p}_0 - p_0)^2] &\leq \frac{2p_0^2(1-p_0)^2}{(N^{(2)})^2} + \frac{p_0(1-p_0)}{(N^{(2)})^3}, \text{ and} \\ \text{Var}_0 \left[ \sum_{i=1}^N (W_i - (N-1)p_0)^2 \right] &= 2N(N-1) [(N-3)p_0^2(1-p_0)^2 + p_0(1-p_0)[p_0^3 + (1-p_0)^3]]. \end{aligned}$$

Hence, we get

$$\text{Var}_0(V) \prec Np_0^2 + p_0 \prec Np_0^2,$$

since  $p_0 \succ 1/N$ . Therefore, by Chebyshev's inequality,  $V = O_P(\sqrt{N}p_0)$ . Under the null,  $N^{(2)}\hat{p}_0 = W \sim \text{Bin}(N^{(2)}, p_0)$ , and because  $N^{(2)}p_0 \rightarrow \infty$ , we have  $\hat{p}_0 \geq \frac{1}{2}p_0$  with probability tending to 1 as  $N \rightarrow \infty$ . We conclude that, under the null,  $V^* = O_P(1)$ .

**Under the alternative.** Turning to the alternative hypothesis, we shall prove that  $V^*$  tends to infinity with high probability by showing that  $\mathbb{E}'_1(V) \gg \sqrt{N}p_0 \vee \sqrt{\text{Var}'_1(V)}$  since  $\sqrt{N}\hat{p}_0 = O_{\mathbb{P}'_1}(\sqrt{N}p_0)$ . The expression (75) of  $V$  still holds.

By definition of  $p_0 = p'_0 + n^{(2)}/N^{(2)}(p_1 - p'_0)$ , we have  $\mathbb{E}'_1(\hat{p}_0) = p_0$ . Furthermore,

$$\begin{aligned} \mathbb{E}'_1[(\hat{p}_0 - p_0)^2] - \frac{1}{N^{(2)}}p_0(1-p_0) &\sim -4\frac{n^2}{N^4}(p_1 - p'_0)^2 \\ \mathbb{E}'_1 \left[ \sum_{i=1}^N (W_i - (N-1)p_0)^2 \right] - N(N-1)p_0(1-p_0) &\sim n^3(p_1 - p'_0)^2 \end{aligned}$$

Inputting this into (75), we get

$$\mathbb{E}'_1[V] \sim (p_1 - p'_0)^2 \frac{n^3}{N}. \quad (76)$$

By (22),  $\mathbb{E}'_1[V] \gg \sqrt{N}p'_0$  and  $\mathbb{E}'_1[V] \gg n^2/N^{3/2}(p_1 - p'_0)$  and it follows that  $\mathbb{E}'_1[V] \gg \sqrt{N}p_0$ . To conclude, we need to control the variance of  $V$  under  $\mathbb{P}'_1$ . Tedious computations lead us to

$$\begin{aligned} \text{Var}'_1[\hat{p}_0 - p_0] &\prec \frac{p_0}{N^2}, \\ \text{Var}'_1[(\hat{p}_0 - p_0)^2] &\prec \frac{p_0^2}{N^4} + \frac{p_0}{N^6}, \text{ and} \\ \text{Var}'_1 \left[ \sum_{i=1}^N (W_i - (N-1)p_0)^2 \right] &\prec N^2p_0 + N^3p_0^2 + n^3(p_1 - p'_0)^2 + n^3Np_0(p_1 - p'_0)^2 + n^4(p_1 - p'_0)^3, \end{aligned}$$

so that, using the fact that  $p_0 \succ 1/N$ , we get

$$\text{Var}'_1[V] \prec Np_0^2 + \frac{n^3}{N}p_0(p_1 - p'_0)^2 + \frac{n^4}{N^2}(p_1 - p'_0)^3.$$

We conclude that  $\mathbb{E}'_1[V] \gg \sqrt{\text{Var}'_1(V)}$  as soon as the following conditions are met

$$(p_1 - p'_0)^2 \frac{n^3}{N} \gg \sqrt{N}p_0, \quad (p_1 - p'_0) \frac{n^{3/2}}{N^{1/2}} \gg \sqrt{p_0}, \quad n(p_1 - p'_0)^{1/2} \gg 1.$$

We already argued that the first one holds, while the second and third are easily seen to be implied by the first condition and the fact that  $p_0 \succ 1/N$ .

## 7.8 Proof of Proposition 5

It suffices to show that the scan test is asymptotically powerful for  $\hat{H}_0$  versus  $H_1$ , where the model under  $\hat{H}_0$  is  $\mathbb{G}(N, \hat{p}_0)$ . In view of Proposition 3, it is therefore enough to prove that, under (23), we have

$$\liminf \frac{nH_{\hat{p}_0}(p_1)}{2\log(N/n)} > 1.$$

First note that  $\hat{p}_0 = W/N^{(2)}$  is concentrated around its mean. Indeed, we have

$$\mathbb{E}[N^{(2)}\hat{p}_0] = (N^{(2)} - n^{(2)})p_0 + n^{(2)}p_1 = N^{(2)}p_0 + n^{(2)}(p_1 - p_0),$$

and

$$\text{Var}[N^{(2)}\hat{p}_0] = (N^{(2)} - n^{(2)})p_0(1 - p_0) + n^{(2)}p_1(1 - p_1) \leq \mathbb{E}[N^{(2)}\hat{p}_0].$$

Hence, by Chebyshev's inequality,

$$\hat{p}_0 = p_0 + a + O_P\left(\frac{1}{N}\sqrt{p_0 + a}\right), \quad a := \frac{n^{(2)}}{N^{(2)}}(p_1 - p_0).$$

Since  $p_0 \gg N^{-2}$ , we have  $\sqrt{p_0}/N = o(p_0)$ . If  $a \geq p_0$ , then  $p_0 \gg N^{-2}$  implies that  $\sqrt{a}/N = o(a)$ . All in all, we get  $\sqrt{p_0 + a}/N = o(p_0 + a)$  and  $\hat{p}_0 \sim_P p_0 + a$ . As in the previous proofs, we can assume that  $p_1/p_0 \rightarrow r \in [1, \infty]$ . In the three following cases, we prove that

$$\liminf \frac{nH_{\hat{p}_0}(p_1)}{2\log(N/n)} > 1.$$

**CASE 1:**  $p_1/p_0 \rightarrow 1$ . In that case, we have  $a = o(p_0)$  and  $\sqrt{p_0}/N = o(p_1 - p_0)$  since

$$\frac{(p_1 - p_0)^2}{p_0} \succ H_{p_0}(p_1) \succ \frac{\log(N/n)}{n}.$$

Hence,  $\hat{p}_0 - p_0 = o(p_1 - p_0)$  and we conclude that

$$H_{\hat{p}_0}(p_1) \sim_P \frac{(p_1 - \hat{p}_0)^2}{2p_0(1 - p_0)} \geq \frac{(p_1 - p_0)^2 - 2(p_1 - p_0)(\hat{p}_0 - p_0)}{2p_0(1 - p_0)} \sim_P H_{p_0}(p_1).$$

**CASE 2:**  $p_1/p_0 \rightarrow r \in (1, \infty)$ . Hence,  $a = o(p_0)$  and  $p_0 \sim_P \hat{p}_0$ . It follows that

$$H_{\hat{p}_0}(p_1) \sim_P \hat{p}_0(r \log(r) - r + 1) \sim_P H_{p_0}(p_1).$$

**CASE 3:**  $p_1/p_0 \rightarrow \infty$ . Since  $\hat{p}_0 \sim_P p_0 + a$ , we derive

$$H_{\hat{p}_0}(p_1) \sim_P p_1 \log \left( \frac{p_1}{p_0 + a} \right) \geq p_1 \log \left( \frac{p_1}{2(p_0 \vee a)} \right) \sim_P H_{p_0}(p_1) \wedge 2p_1 \log \left( \frac{N}{n} \right)$$

It remains to prove that  $\liminf np_1 > 1$  when  $\liminf nH(p_1)/\log(N/n) > 2$ . Assume that  $\liminf np_1 \leq 1$  so that there exists a subsequence satisfying

$$\lim np_1 \leq 1 \text{ and } \liminf np_1 \frac{\log(p_1/p_0)}{\log(N/n)} > 2.$$

It follows that  $\liminf \log(1/(np_0))/\log(N/n) > 2$  and  $\limsup N^2 p_0/n \leq 1$ , which contradicts the assumption of the proposition.

## 7.9 Proof of Proposition 6

We prove the result when  $p_0$  is known. The situation when  $p_0$  is unknown can be dealt with in a similar way; see, for example, the proof of Proposition 5. Let  $\mathbf{B} = \mathbf{W}^2$ . We first lower bound  $\text{SDP}_n(\mathbf{W}^2)$  from below under the alternative where  $S$  is the anomalous subset of indices. We have

$$\text{SDP}_n(\mathbf{B}) \geq \lambda_n^{\max}(\mathbf{B}) \geq \lambda_n^{\max}(\mathbf{B}_S) \geq \frac{1}{n} \mathbf{1}_S^\top \mathbf{W}^2 \mathbf{1}_S = \frac{1}{n} \sum_{i,j \in S} \sum_{k=1}^N W_{ik} W_{kj}.$$

We have

$$\begin{aligned} \mu_S &:= \frac{1}{n} \mathbb{E}_S \left( \sum_{i,j \in S} \sum_{k=1}^N W_{ik} W_{kj} \right) \\ &= [(n-1)p_1 + (N-n)p_0] + (n-1)[(n-2)p_1^2 + (N-n)p_0^2], \\ &= (N-1)p_0 + (n-1)(p_1 - p_0) + (n-1)(N-2)p_0^2 + (n-1)(n-2)(p_1^2 - p_0^2), \end{aligned}$$

and, after some tedious but straightforward calculations,

$$\sigma_S^2 := \frac{1}{n^2} \text{Var}_S \left( \sum_{i,j \in S} \sum_{k=1}^N W_{ik} W_{kj} \right) = O \left( (N/n)p_0(1-p_0)[1 + (np_0)^2] + p_1(1-p_1)[1 + (np_1)^2] \right).$$

By Chebyshev's inequality, under the alternative,  $\text{SDP}_n(\mathbf{B}) \geq \mu_S - O_P(\sigma_S)$ .

Under the null, we bound  $\text{SDP}_n(\mathbf{B})$  from above as [Berthet and Rigollet \(2012\)](#) do. Specifically, they use a result of [Bach et al. \(2010\)](#), which says that

$$\text{SDP}_n(\mathbf{B}) = \min_{\mathbf{U}} \lambda^{\max}(\mathbf{B} + \mathbf{U}) + n|\mathbf{U}|_\infty,$$

where the minimum is over symmetric matrices  $\mathbf{U} = (U_{ij})$  and  $|\mathbf{U}|_\infty := \max_{i,j} |U_{ij}|$ . Similar to what [Berthet and Rigollet \(2012\)](#) do, we apply this identity to  $\mathbf{U} = (U_{ij})$  with  $U_{ij} = -B_{ij} \mathbb{1}_{\{|B_{ij}| \leq z\}}$ , obtaining

$$\text{SDP}_n(\mathbf{B}) \leq \lambda^{\max}(\tau_z(\mathbf{B})) + nz.$$

where  $\tau_z(\mathbf{B})$  is the hard thresholding of  $\mathbf{B}$  at threshold  $z$ , meaning the matrix with  $(i, j)$  coefficient equal to  $B_{ij} \mathbb{1}_{\{|B_{ij}| > z\}}$ . Under the null, we have

$$B_{ii} = \sum_k W_{ik} \sim \text{Bin}(N-1, p_0),$$

and, when  $i \neq j$ ,

$$B_{ij} = \sum_k W_{ik} W_{jk} \sim \text{Bin}(N-2, p_0^2).$$

Fix  $\varepsilon > 0$ . Using Bernstein's inequality (Lemma 2) and the union bound, we find that the following inequalities happen simultaneously with probability tending to one under the null:

$$\max_i B_{ii} \leq (N-1)p_0 + x_0, \quad x_0 := \sqrt{(1+\varepsilon)2Np_0(1-p_0)\log N} + (1+\varepsilon)\log(N),$$

$$\max_{i \neq j} B_{ij} \leq (N-2)p_0^2 + x_{00}, \quad x_{00} := 2\sqrt{(1+\varepsilon)Np_0^2(1-p_0^2)\log N} + 2(1+\varepsilon)\log(N).$$

Hence, choosing  $z = (N-2)p_0^2 + x_{00}$ , we have

$$\text{SDP}_n(\mathbf{B}) \leq \zeta := (N-1)p_0 + x_0 + n(N-2)p_0^2 + nx_{00},$$

with high probability under the null. In order to conclude, we need to prove that  $\mu_S - O(\sigma_S) > \zeta$  with probability going to one.

Before proceeding, we note that (25) implies that, for some  $\eta > 0$ ,

$$np_1^2 > 2(1+\eta)p_0\sqrt{N\log N},$$

and (1) implies that either  $np_0 \geq 1$ , or  $(N/n)^{-a} < np_0 < 1$ , for some sequence  $a \rightarrow 0$ . In particular, this implies

$$n^2 \geq np_0\sqrt{N\log(N)} \geq \sqrt{N}(n/N)^a,$$

so that  $n \geq N^{1/4-a}$ . It also follows that  $n\sqrt{p_0} > \sqrt{n}(N/n)^{-a} \rightarrow \infty$ .

We have  $\mu_S - \zeta \geq (1+o(1))n^2p_1^2 - Np_0^2 - x_0 - nx_{00}$ , with

$$\frac{n^2p_1^2}{Np_0^2} > \frac{2(1+\eta)np_0\sqrt{N\log N}}{Np_0^2} \asymp \frac{n\sqrt{\log N}}{\sqrt{N}p_0} > \frac{n^2\sqrt{\log N}(N/n)^a}{\sqrt{N}} \geq \sqrt{\log N} \rightarrow \infty,$$

$$\frac{x_0}{nx_{00}} \leq \frac{1}{n\sqrt{p_0}} + \frac{1}{n} \rightarrow 0,$$

$$\frac{nx_{00}}{n^2p_1^2} \leq \frac{2n\sqrt{(1+\varepsilon)Np_0^2\log N} + 2(1+\varepsilon)n\log(N)}{2(1+\eta)np_0\sqrt{N\log N}} = \frac{1+\varepsilon}{1+\eta} + \sqrt{\frac{(1+\varepsilon)\log N}{(1+\eta)Np_0^2}} = \frac{1+\varepsilon}{1+\eta} + o(1),$$

since  $Np_0^2 > Nn^{-2}(N/n)^{-2a} > N^{2t-2a}$  with  $2t-2a \rightarrow 2t > 0$ . Assuming that  $\eta > \varepsilon$ , it remains to show that  $n^2p_1^2 \gg \sigma_S$  to prove that  $\mu_S - O(\sigma_S) > \zeta$  with probability going to one in the asymptote.

We have  $\sigma_S^2 \asymp Np_0/n + Nnp_0^3 + n^2p_1^3$ , and

$$\frac{n^2p_1^2}{\sqrt{Np_0/n}} \asymp n\sqrt{p_0}\sqrt{n\log N} \rightarrow \infty,$$

since  $n\sqrt{p_0} \rightarrow \infty$ , and also

$$\frac{n^2p_1^2}{\sqrt{Nnp_0^3}} \asymp \sqrt{n\log(N)/p_0} \rightarrow \infty,$$

and

$$\frac{n^2p_1^2}{\sqrt{n^2p_1^3}} = n\sqrt{p_1} \geq np_1 \rightarrow \infty.$$

### 7.10 Proof of Proposition 7

The first results follows from a simple consequence of Bernstein's inequality for binomial random variables and the union bound. Details are omitted. Let us concentrate on the second bound. It suffices to prove that with probability  $\mathbb{P}_S$  going to one  $\max_{i \in S} W_i. < \max_{i \in S^c} W_i.$  since the distribution of  $\max_{i \in S^c} W_i.$  under  $\mathbb{P}_S$  is stochastically smaller than the distribution of  $\max_{i=1, \dots, N} W_i.$  under  $\mathbb{P}_0.$  Since  $\limsup \log(n)/\log(N) < 1$ , we can assume that  $n < N^{1-\epsilon}$  for some  $\epsilon > 0$ . Condition (1) also enforces  $p_0 \gg \log(N)/N$ . Since the power of the maximal degree test is increasing with respect to  $p_1$ , we can assume that  $p_1$  satisfies Condition (27) but is still large enough so that  $p_1 \gg \log(n)/n$ .

Fix  $\delta > 0$  arbitrarily small. Applying Bernstein's inequality (Lemma 2) and using  $Np_0 \gg \log(N)$  and  $np_1 \gg \log(n)$ , we derive that

$$\begin{aligned} \max_{i \in S} W_i. & - (N-1)p_0 + (n-1)(p_1 - p_0) \\ & \leq \sqrt{2(1+\delta)(N-1)p_0(1-p_0)\log(n)} + \sqrt{(2+\delta)np_1(1-p_1)\log(n)} \\ & \leq \sqrt{2(1+\delta)(1-\epsilon)(N-1)p_0(1-p_0)\log(n)}(1+o(1)) . \end{aligned} \quad (77)$$

with probability going to one since we assume that  $n(p_1 - p_0) = o(\sqrt{N \log(N) p_0}) = o(Np_0)$ .

Let us consider a subset  $T \subset S^c$  of size  $N^{1-\kappa}$  with some  $\kappa > 0$ . As the  $W_i.$  are not independent, it is not straightforward to directly lower bound their supremum. This is why we compare it to independent variables. Let us call the  $i_T^*$  the smallest  $i$  in  $T$  that achieves  $\max_{i \in T} \sum_{j \in T^c} W_{i,j}$

$$\max_{i \in S^c} W_i. \geq \max_{i \in T} W_i. \geq \max_{i \in T} \sum_{j \in T^c} W_{i,j} + \sum_{j \in T} W_{i_T^*,j} .$$

Observe that the first term is supremum of  $|T|$  independent binomial variables and that the second term follows a binomial distribution with parameters  $p$  and  $|T|$ . With probability going to one, we have  $\sum_{j \in T} W_{i_T^*,j} \geq |T|p_0 - \sqrt{|T|p_0(1-p_0)\log(|T|)}$ . Let us turn to the supremum of independent binomial distributions. We start from  $\mathbb{P}(\text{Bin}(n, p) = k) = p^k(1-p)^{n-k} \binom{n}{k}$ . Consider  $p$  bounded away from 1 and  $k \geq np$  such that  $k/n$  is also bounded away from one. Using the stirling formula  $\sqrt{2\pi n}(n/e)^n < n! < \sqrt{2\pi n}(n/e)^n e^{1/(12n)}$ , we get

$$\begin{aligned} \mathbb{P}(\text{Bin}(n, p) = k+i) & \geq \exp[-nH_p(k/n)] \frac{1}{e^2 \sqrt{2\pi k}} \left( \frac{p(1-k/n)}{k/n(1-p)} \right)^i \left( 1 - \frac{i}{k+i} \right)^i , \\ \mathbb{P}(\text{Bin}(n, p) \geq k) & \succ \exp[-nH_p(k/n)] \frac{1}{\sqrt{k}} \frac{k/n(1-p)}{k/n-p} \left[ 1 - \left( \frac{p(1-k/n)}{k/n(1-p)} \right)^{\sqrt{k}} \right] , \end{aligned}$$

where we have summed the first inequality for  $i = 0, \dots, \sqrt{k} - 1$ . Applying this lower bound to  $\sum_{j \in T^c} W_{i,j}$  and using Lemma 3, we derive that

$$\mathbb{P}_S \left[ \sum_{j \in T^c} W_{i,j} \geq (N-1-|T|)p_0 + \sqrt{2(1-\delta)(N-|T|-1)p_0(1-p_0)\log(|T|)} \right] \succ \frac{|T|^{1-\delta}}{\sqrt{(1-\delta)\log(|T|)}} .$$

Since the random variables  $\sum_{j \in T^c} W_{i,j}$  for  $i \in T$  are independent, it follows that

$$\sup_{i \in T} \sum_{j \in T^c} W_{i,j} \geq (N-1-|T|)p_0 + \sqrt{2(1-\delta)(N-|T|-1)p_0(1-p_0)\log(|T|)} ,$$



with probability going to one. All in all, we derive that with probability going to one

$$\max_{i \in S^c} W_i \geq (N-1)p_0 + \sqrt{2(1-\delta)(N-1)p_0(1-p_0)(1-\kappa)\log(N)} - 2\sqrt{2N^{1-\kappa}p_0(1-p_0)\log(N)},$$

where the last term is negligible in front of the second term. Comparing this last lower bound with (77) and taking  $\kappa$  and  $\delta$  small enough allows us to conclude.

### 7.11 Proof of Proposition 8

By Chebyshev's inequality  $h(\mathcal{V}) \sim_{\mathbb{P}_0} Np_0/2$  with probability going to one. Since  $h(S) \leq |S|/2$ , we have  $h(S) \leq Np_0/4$  for all subsets  $S$  of size smaller than  $Np_0/2 \rightarrow \infty$ . Note that  $|S|h(S) \sim \text{Bin}(|S|^{(2)}, p_0)$ . Applying Bernstein inequality (Lemma 2) and Lemma 4 to all subsets  $S$  of size larger than  $Np_0/2$ , we derive that

$$|S|h(S) \leq \frac{|S|^2}{2}p_0 + \sqrt{|S|^3p_0(1-p_0)\log\left(\frac{Ne^2}{|S|}\right)} + |S|\log\left(\frac{Ne^2}{|S|}\right)$$

with probability larger than  $1 - \exp(-Np_0/2)$ . Comparing  $h(S)$  with  $Np_0/2$ , we get

$$\frac{2h(S)}{Np_0} \leq \frac{|S|}{N} + 2\sqrt{\frac{|S|}{N}}\sqrt{\frac{2 + \log(N/|S|)}{Np_0}} + 2\frac{2 + \log(N/|S|)}{Np_0} = \frac{|S|}{N} + 2\sqrt{\frac{|S|}{N}}o(1) + o(1),$$

since  $Np_0 \gg \log(N)$ . This quantity is away from one, except if  $|S| \sim N$ . As a consequence,  $\max_S h(S) \sim_{\mathbb{P}_0} h(\mathcal{V}) \sim_{\mathbb{P}_0} Np_0/2$  with probability going to one.

Let us turn to the alternative distribution. Under  $\mathbb{P}_S$ ,  $|S|h(S) \sim \text{Bin}(|S|^{(2)}, p_1)$ . It follows that  $h(S) \sim_{\mathbb{P}_S} np_1/2$  with probability going to one. The densest subgraph test is therefore powerful when  $\liminf \frac{np_1}{Np_0} > 1$ .

Let us now assume that  $\frac{np_1}{Np_0} \rightarrow 0$ . For any subset  $T$ ,  $|T|h(T)$  is the sum of two independent binomial distributions of parameters  $(|S \cap T|^{(2)}, p_1)$  and  $(|T|^{(2)} - |S \cap T|^{(2)}, p_0)$ . Applying, as previously, Bernstein's inequality for all subsets  $T$  of size larger than  $Np_0/2$ , we derive that

$$\begin{aligned} |T|h(T) &\leq \frac{|T|^2p_0}{2} + \frac{|S \cap T|^2}{2}(p_1 - p_0) + \sqrt{|T|^3p_0\log\left(\frac{Ne^2}{|T|}\right)} + |T|\log\left(\frac{Ne^2}{|T|}\right) \\ &\quad + \sqrt{2|T \cap S|^3p_1\log\left(\frac{ne}{|S \cap T| \vee 1}\right)} + 2|S \cap T|\log\left(\frac{ne}{|S \cap T| \vee 1}\right) \end{aligned}$$

with probability going to one. Comparing  $h(T)$  with  $Np_0/2$  we get

$$\frac{2h(T)}{Np_0} \leq \frac{|T|}{N} + \frac{n}{N}\frac{p_1 - p_0}{p_0} + o(1) + \sqrt{\frac{np_1}{Np_0}}o(1).$$

Since we assume that  $np_1 = o(Np_0)$ , this quantity is away from one except if  $|T| \sim N$ .

## Acknowledgements

We would like to thank Jacques Verstraete for helpful discussions on the clique number of a random graph. We first learned about the work of Butucea and Ingster (2011) at the *New Trends in Mathematical Statistics* conference held at the Centre International de Rencontres Mathématiques (CIRM), Luminy, France, in 2011. The research of E. Arias-Castro is partially supported by a grant from the Office of Naval Research (N00014-13-1-0257). The research of N. Verzelen is partly supported by the french Agence Nationale de la Recherche (ANR 2011 BS01 010 01 projet Calibration).

## References

- Albert, R. and A.-L. Barabási (2002, Jan). Statistical mechanics of complex networks. *Rev. Mod. Phys.* 74, 47–97.
- Aldous, D. J. (1985). *Exchangeability and related topics*, École d’été de probabilités de Saint Flour XIII, Volume 1117 of *Lecture Notes in Mathematics*. Berlin: Springer-Verlag.
- Alon, N., M. Krivelevich, and B. Sudakov (1998). Finding a large hidden clique in a random graph. In *Proceedings of the Eighth International Conference “Random Structures and Algorithms” (Poznan, 1997)*, Volume 13, pp. 457–466.
- Arias-Castro, E., E. J. Candès, and Y. Plan (2011). Global testing under sparse alternatives: Anova, multiple comparisons and the higher criticism. *Ann. Statist.* 39(5), 2533–2556.
- Bach, F., S. Ahipasaoglu, and A. d’Aspremont (2010). Convex relaxations for subset selection. *arXiv preprint arXiv:1006.3601*.
- Barabási, A.-L. and R. Albert (1999). Emergence of scaling in random networks. *Science* 286(5439), 509–512.
- Berthet, Q. and P. Rigollet (2012). Optimal detection of sparse principal components in high dimension. Available online at <http://arXiv.org/abs/1202.5070>.
- Bickel, P. J. and A. Chen (2009). A nonparametric view of network models and newman–girvan and other modularities. *Proceedings of the National Academy of Sciences* 106(50), 21068–21073.
- Bickel, P. J., A. Chen, and E. Levina (2011). The method of moments and degree distributions for network models. *Ann. Statist.* 39(5), 2280–2301.
- Bollobás, B. (2001). *Random graphs* (Second ed.), Volume 73 of *Cambridge Studies in Advanced Mathematics*. Cambridge: Cambridge University Press.
- Butucea, C. and Y. I. Ingster (2011). Detection of a sparse submatrix of a high-dimensional noisy matrix. Available from <http://arxiv.org/abs/1109.0898>.
- d’Aspremont, A., L. El Ghaoui, M. I. Jordan, and G. R. G. Lanckriet (2007). A direct formulation for sparse pca using semidefinite programming. *SIAM Review* 49(3), 434–448.
- Dekel, Y., O. Gurel-Gurevich, and Y. Peres (2011). Finding hidden cliques in linear time with high probability. In *Proceedings of the Eighth Workshop on Analytic Algorithmics and Combinatorics (ANALCO)*, pp. 67–75. Society for Industrial and Applied Mathematics (SIAM).
- Donoho, D. and J. Jin (2004). Higher criticism for detecting sparse heterogeneous mixtures. *Ann. Statist.* 32(3), 962–994.
- Feige, U. and D. Ron (2010). Finding hidden cliques in linear time. In *21st International Meeting on Probabilistic, Combinatorial, and Asymptotic Methods in the Analysis of Algorithms (AofA’10)*, Discrete Math. Theor. Comput. Sci. Proc., AM, pp. 189–203. Assoc. Discrete Math. Theor. Comput. Sci., Nancy.
- Feldman, V., E. Grigorescu, L. Reyzin, S. Vempala, and Y. Xiao (2012). Statistical algorithms and a lower bound for planted clique. Available online at <http://arXiv.org/abs/1201.1214>.

- Fortunato, S. (2010). Community detection in graphs. *Physics Reports* 486(3–5), 75 – 174.
- Girvan, M. and M. E. J. Newman (2002). Community structure in social and biological networks. *Proceedings of the National Academy of Sciences* 99(12), 7821–7826.
- Hall, P. and J. Jin (2010). Innovated higher criticism for detecting sparse signals in correlated noise. *Ann. Statist.* 38(3), 1686–1732.
- Ingster, Y., A. Tsybakov, and N. Verzelen (2010). Detection boundary in sparse regression. *Electronic Journal of Statistics* 4, 1476–1526.
- Ingster, Y. I. (1997). Some problems of hypothesis testing leading to infinitely divisible distributions. *Math. Methods Statist.* 6(1), 47–69.
- Ingster, Y. I. and I. A. Suslina (2002). On the detection of a signal with a known shape in a multichannel system. *Zap. Nauchn. Sem. S.-Peterburg. Otdel. Mat. Inst. Steklov. (POMI)* 294 (Veroyatn. i Stat. 5), 88–112, 261.
- Karp, R. M. (1972). Reducibility among combinatorial problems. In *Complexity of computer computations (Proc. Sympos., IBM Thomas J. Watson Res. Center, Yorktown Heights, N.Y., 1972)*, pp. 85–103. New York: Plenum.
- Khuller, S. and B. Saha (2009). On finding dense subgraphs. *Automata, Languages and Programming*, 597–608.
- Kulldorff, M. (1997). A spatial scan statistic. *Comm. Statist. Theory Methods* 26(6), 1481–1496.
- Lancichinetti, A. and S. Fortunato (2009, Nov). Community detection algorithms: A comparative analysis. *Phys. Rev. E* 80, 056117.
- Lehmann, E. L. and J. P. Romano (2005). *Testing statistical hypotheses* (Third ed.). Springer Texts in Statistics. New York: Springer.
- Newman, M. E. J. (2006). Modularity and community structure in networks. *Proceedings of the National Academy of Sciences* 103(23), 8577–8582.
- Newman, M. E. J. and M. Girvan (2004, Feb). Finding and evaluating community structure in networks. *Phys. Rev. E* 69, 026113.
- Reichardt, J. and S. Bornholdt (2006, Jul). Statistical mechanics of community detection. *Phys. Rev. E* 74, 016110.
- Rossman, B. (2010). *Average-Case Complexity of Detecting Cliques*. Ph. D. thesis, Massachusetts Institute of Technology.
- Zuckerman, D. (2006). Linear degree extractors and the inapproximability of max clique and chromatic number. In *STOC’06: Proceedings of the 38th Annual ACM Symposium on Theory of Computing*, pp. 681–690. New York: ACM.